IDENTIFICATION OF BODY FLUIDS BY mRNA ANALYSIS WITH MINION NANOPORE SEQUENCING

A thesis presented to the faculty of the Graduate School of Western Carolina University in partial fulfillment of the requirements for the degree of Master of Science in Biology

By

Mary Jessamine Michaels

Director: Dr. Kelly Grisedale
Assistant Professor, Department of Chemistry and Physics
Forensic Science Program Coordinator

Committee Members: Brittania Bintz, Department of Chemistry and Physics
Dr. Maria Gainey, Department of Biology

May 2018

ACKNOWLEDGMENTS

I would like to express my sincerest gratitude to my thesis advisor, Kelly Grisedale, and to Brittania Bintz of WCU's Forensic Science Department, for all of their support throughout this project. They have both made themselves available to me for both personal and research advice and guidance, and this project would not have been possible without them. I would also like to thank my other committee member, Maria Gainey, for her invaluable insight and aid in the successful completion of this project.

I would like to acknowledge the WCU's Biology Department for allowing me to continue my education and providing me with this incredible experience. I would also like to thank WCU's Forensic Science Department for the providing the financial support and resources needed for this project. Additionally, I would like to acknowledge the wonderful donors who contributed their fluids for use in this project, as none of this would have been possible without them.

Lastly, I would like to thank all the friends, family, and colleagues who have offered their support throughout the length of this project. I would like to thank my boyfriend, Keaton Foster, for his support and endless help with coding and data analysis. I would especially like to thank my parents and grandparents for their ongoing support, as none of this would have been possible without their help and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

18S rRNA—18S ribosomal RNA
ALAS2—δ-Aminolevulinate synthase
B2M—beta-2-microglobin
Bp—base pairs
CDK—cyclin dependent kinase
cDNA—complementary DNA
CYP2B7P1—cytochrome p450, family 2, subfamily B, polypeptide 7, pseudogene 1
DNA—deoxyribonucleic acid
dNTPs—deoxynucleotide
DTT—Dithiothreitol
EDNAP—European DNA Profiling Group
G6PDH—Glucose 6-phosphate dehydrogenase
gDNA—genomic DNA
HBA—hemoglobin A
HBB—hemoglobin B
HBD1—human beta-defensin 1
HTN3—histatin 3
IDT—Integrated DNA Technologies
LLB—library loading beads
miRNA—micro ribonucleic acid
MMLV—Moloney murine leukemia virus
MMP7—matrix metalloproteinase 7
MMP10—matrix metalloproteinase 10
MMP11—matrix metalloproteinase 11
mRNA—messenger ribonucleic acid
MPS—massively parallel sequencing
MUC2—mucin 4
MUC7—mucin 7
MYOZ1—myozenin 1
NADPH—nicotinamide adenine dinucleotide phosphate
NEB—New England Biosystems
ONT—Oxford Nanopore Technologies
PCR—polymerase chain reaction
PRM1—protamine 1
PRM2—protamine 2
RBF—running buffer with fuel mix
RIN—RNA integrity number
RNA—ribonucleic acid
SEMG1—seminogelin 1
STATH—statherin
STR—short tandem repeats
Taq—polymerase from *Thermus aquaticus*

TS—template switching
UBC—Ubiquitin C

ABSTRACT

IDENTIFICATION OF BODY FLUIDS BY MRNA ANALYSIS WITH MINION NANOPORE SEQUENCING

Mary Jessamine Michaels

Western Carolina University (May 2018)

Director: Kelly Grisedale, PhD.

The identification of body fluids present on evidence items in a criminal investigation can be vital to understanding the nature of a crime, particularly in cases of sexual assault. Although crime labs can confirm the presence of body fluids like semen and blood on a piece of evidence using traditional serological techniques, they cannot confirm the presence of saliva or vaginal fluid or differentiate peripheral blood from menstrual blood. Due to the unique patterns of gene expression in different cell types, different body fluids contain distinct messenger RNA (mRNA) molecules, which can be analyzed to generate mRNA profiles for confirmatory identification of body fluids. High throughput sequencing methods present the opportunity to generate large amounts of data from low level forensic samples that is not achievable with more traditional techniques. In this study, the MinION sequencer by Oxford Nanopore Technologies (ONT), which is a small and affordable high throughput sequencer that generates data rapidly and in real time, was assessed to determine its ability to generate high quality data from forensic type samples. Semen, saliva, blood, vaginal fluid, and menstrual blood from eight different donors were collected on sterile swabs. Additionally, a tenfold dilution series was performed on semen, blood, and saliva samples from three donors each and decreasing volumes of semen were

pipetted onto vaginal swabs from three sets of donors. DNA and RNA were co-extracted from half swabs. DNA fractions were taken forward for short tandem repeat (STR) analysis. For RNA fractions, cDNA was generated, and a multiplex PCR targeting two genes per body fluid was performed. Amplicons were sequenced on the MinION with the 1D Ligation Sequencing Kit by ONT. Full DNA profiles were obtained for all semen, blood, and menstrual blood samples, and most of the saliva and vaginal fluid samples. Profiles were obtained for diluted semen, saliva, and blood samples, and mixed profiles were obtained for the semen/vaginal fluid mixture samples. The multiplex PCR was highly specific for each body fluid, with little to no cross reactivity. The MinION was able to obtain at least 1,000X coverage of target genes and little to no off target reads, even for some diluted samples. Both vaginal fluid and semen genes were detected in mixture samples. Optimization of the MinION sequencing workflow to maximize read counts and minimize costs should be explored. Despite the high input requirements stated by ONT for sequencing, the MinION appears to be able to generate high quality data with lower DNA/RNA input.

CHAPTER ONE: INTRODUCTION

Human body fluids discovered at a crime scene can be used as evidence to identify victims or suspects, or to link victims and suspects to each other or to specific crime scenes. When crime labs receive items of evidence, they first analyze the evidence to determine what body fluids are present, which helps to give context to the evidence items and may assist in reconstructing the crime event in question. Once serological testing is completed, evidence items are taken forward for DNA analysis to identify the individual who contributed the body fluid. The combination of serological and DNA analysis results allows investigators to link a particular individual to a crime scene and presents information that gives them a better idea of the circumstances surrounding the crime. Although crime labs can confirm the presence of body fluids like semen and blood on an item of evidence using traditional serological techniques, these tests can be laborious and can consume much of the evidence sample. Furthermore, many crime labs are not able to confirm the presence of saliva or vaginal secretions or differentiate peripheral blood from menstrual blood. Identifying these body fluids is important in oral or digital assaults, when the presence or absence of semen is not informative within the context of the assault, or when menstrual blood must be differentiated from blood present due to trauma. Tests for these body fluids are presumptive only, so there is a need in the field of forensics for robust and reliable tests that can confirm the presence of various body fluids that may be present on a sample.

Molecular body fluid identification techniques

The identification of body fluids on items of evidence can be vital to understanding the nature of a crime. Current serological techniques are completed separately and require multiple

cuttings from the evidence item in question, making them time consuming and inefficient. In addition, many forensic laboratories lack tests that are able to confirm the presence of biological fluids like saliva and vaginal secretions.[1] Emerging molecular techniques for body fluid identification allow for the ability to confirmatively identify various body fluids, as well as allow for the potential to identify different types of body fluid with one all-inclusive test. Many of these techniques are founded in differential gene expression levels in body fluid cells. Different cell types have unique patterns of gene expression within the cell, creating distinct profiles for different body fluid types.[2] Specific messenger RNA (mRNA) markers for various body fluids have been identified and can be used for confirmatory body fluid testing. Analysis of microRNAs and DNA methylation patterns have also been used for body fluid identification. MicroRNAs, which are less subject to degradation, are fairly specific for semen and blood, but are less specific for saliva and vaginal secretions, requiring a greater number of markers for positive identification.[3] DNA methylation pattern analysis often involves interpreting the level of methylation, rather than just the presence or absence of a marker, making interpretation less clear.[3] The highly specific and sensitive markers that have been identified for mRNA make interpretation of results easier.

In the past, the use of mRNA for body fluid identification has been questioned due to its unstable nature. RNA is more susceptible to degradation than DNA and is consequently less stable and more difficult to work with.[1] However, over the past decade, many studies have shown that RNA is detectable in forensic type samples, and robust and reliable protocols using various analysis methods, including sequencing, have been developed. Although hot or humid conditions can often be detrimental to the RNA in the sample, mRNA profiles are obtainable from saliva stains for at least 365 days and in vaginal secretions for at least 547 days when

samples were stored at room temperature.[4] Zhao et al. (2016) found that the hemoglobin markers HBA and HBB were detectable in 30 and 50 year old blood stains.[5] Even when exposed to harshly dry and humid environments, semen and saliva are detectable for up to 71 weeks.[6] Saliva and vaginal secretion stains show less sustainability due to greater susceptibility to hydrolytic damage over time.[4, 6] Despite this, the RNA found in biological stains appears to be stable enough to obtain profiles for identification when samples are properly stored.

DNA/RNA co-extraction

Performing serological testing as part of the DNA analysis workflow means that two cuttings must be consumed for each item of evidence. One cutting is made for body fluid identification and another is made for DNA profiling, consuming much of what may be a very small sample. Using RNA analysis for body fluid identification offers the possibility of co-extraction of DNA and RNA from the same cutting, making analysis more efficient and less wasteful. Although the quantity and integrity of the RNA and DNA extracted using both organic co-extraction methods and commercial co-extraction kits have been assessed, no method has consistently been shown to be optimized for both DNA and RNA analysis. Many studies have used the AllPrep DNA/RNA Mini Kit (Qiagen) for co-extraction.[7-9] Although this co-extraction kit appears to be sufficient for recovery of DNA, the recovery of RNA is poor in comparison to other methods.[7-9] Grabmuller et al. (2015) found that RNA extraction was most efficient with the NucleoSpin® miRNA Kit (Macherey-Nagel) and the RNeasy® Mini Kit (Qiagen). However, the NucleoSpin® protocol involves more handling of the samples, increasing the potential for cross contamination.[7] Conversely, the RNeasy® kit can be automated, which would allow for additional contamination prevention. In their study, which compared the performance of different RNA/DNA co-extraction methods, Grabmuller et al. (2015) isolated DNA from the

lysate waste after the initial spin in the RNeasy® kit workflow and extracted DNA from it with the Prepfiler® Forensic DNA Extraction Kit (Life Technologies). Although the AllPrep kit had slightly better recovery, a comparable amount of DNA was recovered with the RNeasy® kit. [7]

Reliability of mRNA markers

Gene markers used in multiplex RNA analyses must be both body fluid specific and sensitive enough to detect in the low template samples generally encountered in forensic casework. Additionally, primers for such markers must eliminate the possibility of amplification of genomic DNA, which is done by using primers that either span exon-exon boundaries or span an exon and intron region that would allow DNA amplicons to be eliminated due to size differences. [10] While semen, blood, and menstrual blood markers appear to be more sensitive and specific, some saliva and vaginal secretion markers exhibit non-specific amplification. [11-15] For semen samples, the Seminogelin 1 (SEMG1) and Protamine 1 (PRM1) genes appear to be highly successful. [11-15] The seminogelin gene encodes the main protein present in seminal fluid that entraps spermatozoa cells and is essential for flagellar movement of spermatozoa. [16] Protamine 1 encodes for a protein that forms a complex with protamine 2 and histones in spermatozoa cells enabling packaging of chromatin for epigenetic control of expression. [17] Additionally, SEMG1 has been successful for identifying semen from vasectomized donors, while PRM1 has been successful for identification of semen from fertile donors. [11]

Differentiation between peripheral and menstrual blood can be achieved by targeting matrix metalloproteinase genes (MMP7, MMP10, and MMP11), which are involved in the cleavage of cell surface receptors and breakdown of the extracellular matrix, leading to the dissociation of uterine tissues during menstruation. [18] These have been shown to be highly specific to menstrual blood samples. [10, 19] Among other peripheral blood specific genes, the δ-

Aminolevulinate synthase (ALAS2) and Hemoglobin alpha locus 1 (HBA) genes have been shown to be highly sensitive and specific.[5, 10, 20] ALAS2 is a gene that encodes for an enzyme that catalyzes the synthesis of heme in erythrocytes, and HBA is a subunit of hemoglobin, both of which work together to allow for the transport of oxygen through the bloodstream.[21, 22] Due to the presence of peripheral blood in menstrual blood, both peripheral blood and menstrual blood markers are detectable in these samples. Because hemoglobin genes are highly expressed in blood cells, the HBA gene has been shown to be detectable in aged stains.[10]

Due to the expression of similar genes, like mucins, in both saliva and vaginal secretion samples, cross reactivity has been observed in many studies. Targeting of the human beta-defensin 1 (HBD1) and mucin 4 (MUC4) genes in vaginal secretions have shown cross reactivity with saliva samples.[10, 19] The mucin 7 (MUC7) gene targeted for saliva samples has been shown to be cross reactive with vaginal fluid samples.[12] Despite this, the Histatin 3 (HTN3) and Statherin (STATH) genes appear to be both sensitive and specific markers for saliva samples.[12, 15] Histatin 3 encodes a salivary protein that has antimicrobial properties and interacts with heat shock proteins and cyclin dependent kinase (CDK) inhibitors to promote progression through the cell cycle.[23] Statherin encodes a protein that binds to hydroxyapatite found on tooth enamel to prevent calcium phosphate buildup.[24] Although STATH appears to be a good marker for saliva, it is expressed less than HTN3 and is therefore slightly less sensitive.[15]

The myozenin 1 (MYOZ1) and cytochrome p450, family 2, subfamily B, polypeptide 7, pseudogene 1 (CYP2B7P1) genes have been shown to be more specific for vaginal secretions than HBD1 and MUC4. 10 MYOZ1 encodes an α-actin and γ filament binding protein that functions in sarcomeres and is primarily expressed in skeletal muscle.[25] CYP2B7P1 is a non-coding RNA expressed highly in vaginal fluids and internal organs such as the liver and lungs.[26]

Because the only other samples in which these mRNAs would be highly expressed is in deep tissues or organs, theses markers should be specific for vaginal fluid samples found in a forensic context. Although targeting of Lactobacillus genes in vaginal fluid samples has been successful, there is a greater risk of contamination and non-specific amplification of similar or related microbial species present on other parts of the body.[10] In addition to targeting body fluid specific markers, housekeeping genes should also be targeted for use as a positive control. Glucose 6-phosphate dehydrogenase (G6PDH) is a protein that produces reduced nicotinamide adenine dinucleotide phosphate (NADPH), which is an important electron donor involved in cellular respiration, and 18S rRNA is a part of the ribosomal subunit in eukaryotes.[27] Because both of these are expressed in all tissues, they have been shown to be successful housekeeping targets.[10, 11]

MinION nanopore sequencing

In the past, capillary electrophoresis has primarily been used to analyze target mRNAs. However, due to the high throughput of next-generation sequencing technologies, more studies are using these platforms for mRNA analysis. One fairly new high throughput sequencer is the Oxford Nanopore Technologies (ONT) MinION. The MinION is small, portable, relatively inexpensive in comparison to other next-generation DNA sequencers, and produces data in real-time.[28] Rather than employing the sequencing by synthesis method utilized in other platforms, the MinION uses protein nanopores embedded in a membrane through which an electrical current is applied. An enzyme separates DNA strands and pushes a single strand through the nanopore one base at a time. As the strand is passed through, the MinION detects disruptions in the membrane's electrical current.[29] Because different combinations of nucleotides cause distinctive disruptions in the current, the software can call bases according to those current

changes. Although this new technology is promising, there are some issues. Accuracy of basecalling, particularly for longer reads, has been shown to be less than with other sequencing methods.[28, 29] However, changes in the sequencing chemistry have increased accuracy and precision to up to 99% for genotyping.[29] In addition, most of the sequencing kits for the MinION have a high DNA input requirement, which is not ideal for forensic samples. However, its ability to sufficiently sequence forensic type samples is not fully known.

Oxford Nanopore Technologies offers two methods of sequencing (Ligation sequencing and Rapid sequencing) on the MinION. The Rapid Sequencing Kits are designed only for high molecular weight gDNA, and are therefore not suitable for mRNA or amplicon analysis. The Ligation Sequencing Kits, which are able to sequence gDNA, cDNA, and amplicons offer two methods of sequencing: 1D and $1D^2$. The 1D Ligation Sequencing Kit sequences only one strand of the DNA, while the $1D^2$ Ligation Sequencing Kit allows for sequencing of both the template and complimentary strands, allowing for better accuracy. However, the $1D^2$ Ligation Kit has a longer library preparation time and was more expensive than the 1D Ligation Kit at the start of this study, although the kits are now the same price. The ligation kits library preparation starts by dA-tailing DNA fragments and ligating adapters to ends.[30] The adapters are attached to an enzyme that ratchets the DNA strand through the nanopore one base at a time.[30] A specialized sequence in the adapter prevents activation of the enzyme until contact with a nanopore on the flow cell is achieved.[30] During an adapter clean-up elution step, tethers are added to adapter-ligated ends, which allows fragments to locate to nanopores upon loading the library onto the flow cell.[30] Figure 1 shows the library preparation steps for 1D Ligation Sequencing Kit.

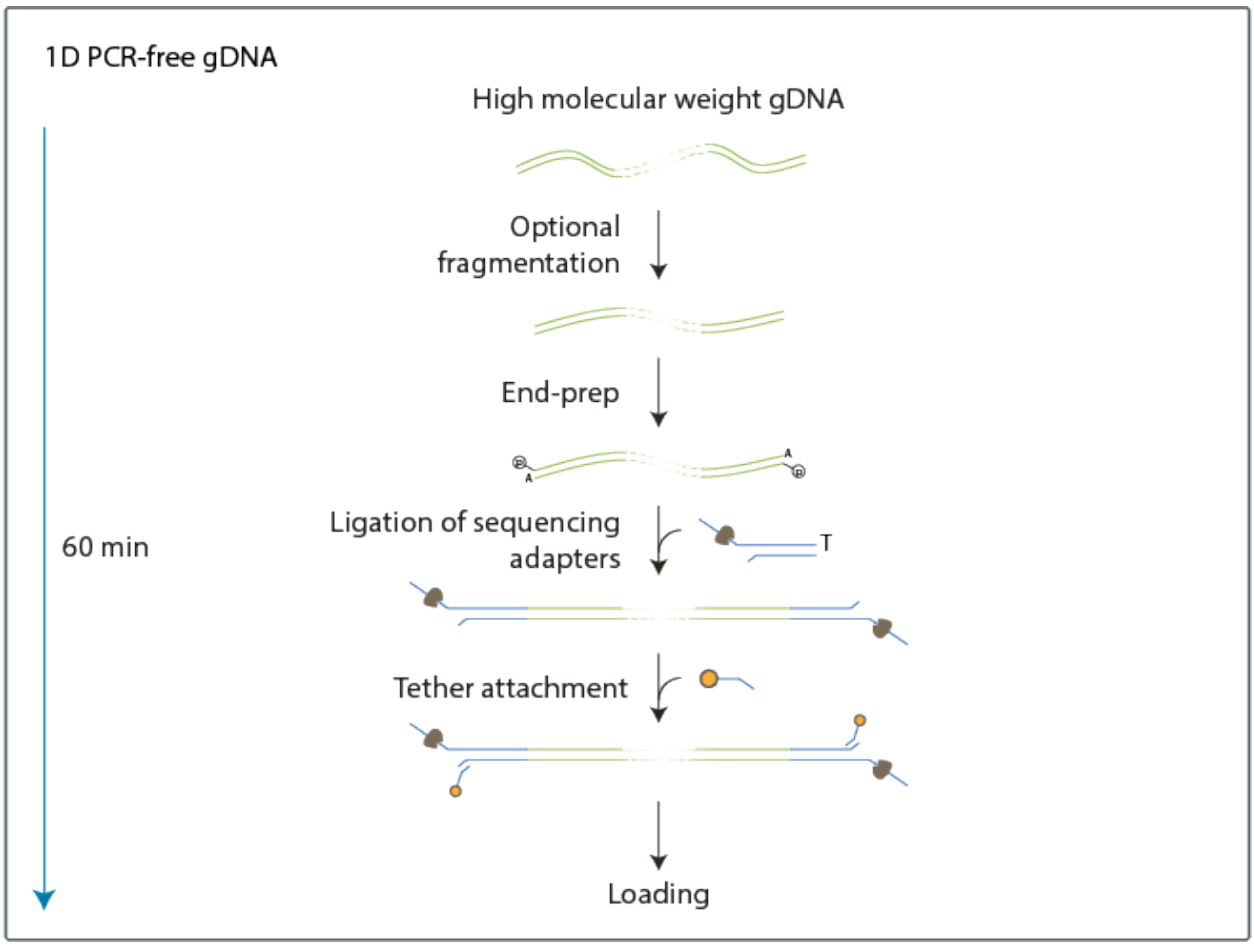

Figure 1. Library preparation for ONT's 1D Ligation Sequencing Kit, consisting of end-prepping of DNA fragments, ligation of adapters to fragment ends, and attachment of tethers to fragment ends. [30]

Purpose of the study

The purpose of this study is to explore a rapid method for identifying body fluids from forensic samples. Because the MinION DNA sequencer is very small, inexpensive, and has a rapid data turnaround in comparison to other high throughput DNA sequencers, it could be ideal for the processing of evidence in crime laboratories. To examine the reliability of this method, semen, saliva, peripheral blood, vaginal fluid, and menstrual blood from 8 different donors per body fluid were analyzed using the MinION sequencer. Dilution series of semen, blood, and saliva, as well as mixtures of semen and vaginal fluid, were also analyzed. A multiplex PCR

targeting two specific genes per body fluid was optimized, including two housekeeping gene targets as endogenous controls. Because the MinION sequencer is a new technology, not much is known about how well it obtains high quality data from forensic samples. This study explored the ability of this new sequencer to work in a forensic setting and determined the robustness, reliability, and reproducibility of the results of analysis of various body fluid samples.

CHAPTER TWO: MATERIALS AND METHODS

Sample Collection

Body fluid samples were collected from 8 different donors per body fluid type. Semen and menstrual blood samples from 8 donors were ordered from Lee BioSolutions™. Saliva, peripheral blood, vaginal fluid, and menstrual blood samples were collected from 8 different donors per body fluid with fully informed consent. Menstrual blood and vaginal fluid samples were collected on sterile swabs by inserting the swab approximately 5 cm (2 inches) into the vaginal canal and rotating approximately 5 times. All donors of vaginal fluid or menstrual blood swabs had not engaged in vaginal sex in the seven days prior to collection. Liquid semen, saliva, peripheral blood, and menstrual blood were pipetted onto sterile swabs in 30 µl aliquots and allowed to dry overnight. Tenfold dilution series for semen, saliva, and blood from 3 donors each were prepared and 30 µl of each diluted sample (1:10, 1:100, and 1:1,000) was pipetted onto sterile swabs. Vaginal fluid and semen mixture samples from 3 sets of donors were set up by pipetting 7.5 µl, 15 µl, and 30 µl of neat semen onto vaginal swabs. Half swabs were used for DNA/RNA co-extraction. Additionally, buccal swabs from each donor were collected to serve as DNA reference samples.

DNA/RNA co-extraction

A DNA/RNA co-extraction method adapted from Grabmuller et al. (2015) was used.[7] The RNeasy® Mini Kit and the Prepfiler® Forensic DNA Extraction Kit were used for co-extraction of DNA and total RNA. 350 µl of Buffer RLT from the RNeasy® kit and 3 µl of Dithiothreitol (DTT) were added to half swabs and incubated at 56ºC for 1 hour. 20 µl of the lysate was removed after incubation and used for extraction of DNA with the Prepfiler® Forensic

DNA Extraction Kit.[31] 300 µl of Prepfiler® Lysis Buffer was added to the DNA fraction without incubation. Then, DNA was extracted from the sample according to the manufacturer's protocol.[31] The remaining lysate was used for RNA extraction with the RNeasy® Mini Kit according to the manufacturer's protocol.[32]

DNA/RNA quantification

DNA samples were quantified using the Quantifiler® Trio DNA Quantification Kit according to the manufacturer's protocol.[33] The concentration of total RNA was analyzed with the Agilent RNA 6000 Pico Kit according to the manufacturer's protocol.[34] An RNA integrity number (RIN) was obtained to assess any sample degradation. An RIN of 10 signifies that the RNA sample is completely intact, while an RIN of 1 signifies that the sample is totally degraded.[35]

DNase treatment and reverse transcription

To eliminate any DNA contamination, total RNA samples were DNase treated using the TURBO DNA-free™ Kit according to the manufacturer's protocol. 1 µl of TURBO™ DNase was added to purified RNA samples with a 0.1 volume of 10X TURBO DNase buffer for routine DNase treatment. After incubation at 37°C for 25 minutes, 2 µl of DNase Inactivation Reagent was added, and samples were incubated at room temperature for 5 minutes before centrifugation.[36] DNase treated samples were then reverse transcribed using the ProtoScript® First Strand cDNA Synthesis Kit according to the manufacturer's protocol.[37]

STR profiling

The GlobalFiler™ PCR Amplification Kit was used for STR analysis of DNA fractions according to the manufacturer's protocol.[38] Amplified samples were then analyzed using the 3500xL Genetic Analyzer, and the data obtained from STR analysis was analyzed using

GeneMapper™ ID-X Software. Profiles obtained from body fluid samples were compared to reference profiles to assess profile completeness. The number of alleles observed for the 21 autosomal markers was determined for each sample, and percent completeness of the profile was calculated.

Multiplex PCR

Short targets

A multiplex PCR was designed to target two genes for each body fluid type. Primer sequences for this multiplex were taken from the literature. Two markers for semen (SEMG1 and PRM1 11), saliva (HTN3 and STATH 39), peripheral blood (HBA 40 and ALAS2 41), vaginal fluid (CYP2B7P1 42 and MYOZ1 41), and menstrual blood (MMP7 and MMP11 11) were included in the multiplex. Additional sequences required for downstream multiplexing with the PCR Barcoding Kit I by Oxford Nanopore Technologies were added to the 5' ends of primer pairs.[43] Primer sequences for all markers are shown in Table 1. Each primer pair was tested with each body fluid type in singleplex reactions to ensure amplification of the proper gene target and to ensure no non-specific amplification of off-target body fluids was observed. For singleplex PCR, 5 units of AmpliTaq Gold™ DNA polymerase (1.0 µl in a 25 µl reaction), 2.5 µl of 10X PCR Buffer, 1.5 µl of 25 mM $MgCl_2$, 2.5 µl of 1.6 µg/µl BSA, 0.5 µl of 10 mM dNTP mix, 1.5 µl of each primer (0.8 µM final concentration), 11.5 µl of RNase-free water, and 2.5 µl of cDNA were combined. The following cycling parameters were used for PCR: initial denaturation at 95°C for 11 minutes and 36 cycles of 95°C for 10 seconds, 63°C for 30 seconds, and 72°C for 30 seconds.

Table 1. Gene target information and primer sequences for multiplex PCR of short amplicons.

	Gene marker	Body fluid	GenBank accession number	Primer sequences (5′ → 3′)	Amplicon Size (bp)	Reference
1	SEMG1	Semen	NM_003007.4	F: GGAAGATGACAGTGATCG T R: CAACTGACACCTTGATATTGG	165	11
2	PRM1	Semen	NM_002761.2	F: AGACAAAGAAGTCGCAGAC T R: TACATCGCGGTCTGTACC	140	11
3	HTN3	Saliva	NM_000200.2	F: GCAAAGAGACATCATGGGTA R:GCCAGTCAAACCTCCATAATC	178	39
4	STATH	Saliva	NM_003154.2	F: TTTGCCTTCATCTTGGCTCT R:CCCATAACCGAATCTTCCAA	138	39
5	ALAS2	Blood	NM_000032.4	F: TTCTGCACCAGAAGGACTCAGCC R:TAAATCTCGCACCCTGGCAGGATC	145	41
6	HBA	Blood	NM_000558.4	F: ACGCTGGCGAGTATGGT R:CCCTTAACCTGGGCAGAG	158	40
7	CYP2B7P1	Vaginal fluid	NR_001278.1	F: AGTCTACCAGGGATATGGCATG R:CTATCAGACACTGAGCCTCGTCC	191	42
8	MYOZ1	Vaginal fluid	NM_021245.3	F: GGGTTGGTGAGACAGGATCA R:TTTTCCCATGGGGAAATATAGGT	129	41
9	MMP7	Menstrual blood	NM_002423.4	F: GAACAGGCTCAGGACTATCTC R:TTTTCCCATGGGGAAATATAGGT	170	11
10	MMP11	Menstrual blood	NM_005940.4	F: CAACCGACAGAAGAGGTTCG R: GAACCGAAGGATCCTGTAGG	120	11
11	18S-rRNA	Housekeeping gene	NR_003286.4	F: CTCAACACGGGAAACCTCAC R: CGCTCCACCAACTAAGAACG	153	11
12	G6PDH	Housekeeping gene	NM_000402.4	F: ATCATCGTGGAGAAGCCCTTC R:GTTCCAGATGGGGCCGA	226	10

*Universal sequences for the ONT barcoding kit were added to the 5′ end of primer sequences
5′ TTTCTGTTGGTGCTGATATTGC–Forward primer 3′
5′ ACTTGCCTGTCGCTCTATCTTC–Reverse primer 3′

A 10X primer mix with 12 primer pairs was prepared containing 2.0 µM of each primer. For the multiplex PCR, 12.5 µl of 2X QIAGEN Multiplex PCR Master Mix, 7.5 µl of RNase-free water, 2.5 µl of 10X primer mix, and 2.5 µl of cDNA was combined for a total reaction volume of 25 µl. Because of unsatisfactory yield of amplification of vaginal fluid and menstrual blood markers, a gradient multiplex PCR was done with vaginal fluid and menstrual blood samples to find the optimal annealing temperature, which was found to be 60°C. The following cycling conditions were used for PCR: initial denaturation at 95°C for 15 minutes, 40 cycles of 95°C for 30 seconds, 60°C for 90 seconds, and 72°C for 90 seconds, and a final extension at 72°C for 10 minutes.[44] In addition, one sample of each body fluid, one of each of the 1:10 dilutions of semen, saliva, and blood, and one of the semen/vaginal fluid mixture samples was sequenced via Sanger sequencing to verify amplicon length observations.

Long targets

Because the length of amplicons needed to obtain high quality sequences with the MinION is generally above 300 base pairs (bp), another multiplex PCR was designed to target longer fragments (≥ 300 bp). All of the same gene targets except for PRM1, MMP11, and the two housekeeping genes, 18S rRNA and G6PDH, were targeted in this multiplex. Because PRM1 is too short to achieve the 300 bp amplicon length, protamine 2 (PRM2) was used to target semen instead of PRM1. Due to the unsuccessful amplification of the markers for menstrual blood in the first multiplex, MMP11 was replaced by another matrix metalloproteinase, MMP10, which has been shown to be more highly expressed in menstrual blood cells than MMP11.[19] Because of the high sensitivity of 18S rRNA marker, and the unsuccessful amplification of G6PDH in many samples, the two housekeeping genes were replaced by ubiquitin C (UBC) and beta-2-microglobin (B2M), which have been successfully used in a series of collaborative efforts by the European DNA profiling group (EDNAP).[12, 19, 45] Primer sequences were designed using the Integrated DNA Technologies® (IDT®) PrimerQuest Tool.[46] Due to the increased susceptibility of mRNA to degradation, the ends of mRNA strands were avoided as targets, so most primers were designed to amplify the innermost area of the mRNA strand. Primers were also designed to span exon-exon junctions to avoid amplification of contaminating DNA. The universal ONT barcoding sequences were added to the 5' ends of primers. Primer sequences used in this multiplex PCR are shown in Table 2.

Table 2. Gene target information and primer sequences for multiplex PCR of long amplicons.

	Gene marker	Body fluid	GenBank accession number	Primer sequences (5′ → 3′)	Amplicon Size (bp)
1	SEMG1	Semen	NM_003007.4	F: CGGAAGAAAGACGACTCCACTA R: ACTGACACCTTGATATTGGTCCAT	428
2	PRM2	Semen	NM_002762.3	F: ATCGCAGAGGCTGCAGAA R: CAAGCTTTATTGGGCAGGTGAC	344
3	HTN3	Saliva	NM_000200.2	F: CTTTCCATGACTGGAGCTGATTCACATGC R: AGGGAAGTATCCTGAAACACAGAATTCAC	344
4	STATH	Saliva	NM_003154.2	F: TGGAAGATTCGGTTATGGGTATG R: GAAACAGTACTGAACACAGCTTAAC	389
5	ALAS2	Blood	NM_000032.4	F: TGCCAGGGTGCGAGATTTA R: GCCAAGAGTTCCAGAGATGATG	397
6	HBA	Blood	NM_000558.4	F: ACAGACTCAGAGAGAACCCA R: CAGTGGCTTAGGAGCTTGAA	380
7	CYP2B7P1	Vaginal fluid	NR_001278.1	F: CTCTATATCCAGCCAGCTGTT R: CTCTGCGACATGAGGGTATTT	381
8	MYOZ1	Vaginal fluid	NM_021245.3	F: GATGTGATGTTGGAGGAACTG R: CTGCCTGGTCTCCTGAT	432
9	MMP7	Menstrual blood	NM_002423.4	F: GGTCACCTACAGGATCGTATCA R: GAGGAATGTCCCATACCCAAAG	398
10	MMP10	Menstrual blood	NM_002425.2	F: ACCCACCTTACATACAGGATTG R:TGAGCAGCAACGAGGAATAA	364
11	B2M	Housekeeping gene	NM_004048.2	F: GGCATTCCTGAAGCTGACA R: CTGCTTACATGTCTCGATCCC	440
12	UBC	Housekeeping gene	NM_021009.6	F: TCGTCACTTGACAATGCAGAT R: CCTTCCTTGTCCTGGATCTTTG	388

*Universal sequences for the ONT barcoding kit were added to the 5′ end of primer sequences
5′ TTTCTGTTGGTGCTGATATTGC–Forward primer 3′
5′ ACTTGCCTGTCGCTCTATCTTC–Reverse primer 3′

All primers were tested in singleplex reactions to ensure proper gene target amplification and to ensure there is no non-specific amplification of off-target body fluids. For the singleplex PCR, 5 units of AmpliTaq Gold™ DNA polymerase (1.0 µl in a 25 µl reaction), 2.5 µl of 10X PCR Buffer, 1.5 µl of 25 mM $MgCl_2$, 2.5 µl of 1.6 µg/µl BSA, 0.5 µl of 10 mM dNTP mix, 1.5 µl of each primer (0.8 µM final concentration), 11.5 µl of RNase-free water, and 2.5 µl of cDNA were combined. For the multiplex, 10X primer mix with all 12 primer pairs was prepared containing 2.0 µM of each primer. The multiplex PCR was prepared with 12.5 µl of 2X QIAGEN Multiplex PCR Master Mix, 7.5 µl of RNase-free water, 2.5 µl of 10X primer mix, and 2.5 µl of cDNA was combined for a total reaction volume of 25 µl. A gradient multiplex PCR was performed to find the best annealing temperature for primers in the multiplex, which was found to be 62°C. The following cycling conditions were used for PCR: initial denaturation at

95°C for 15 minutes, 40 cycles of 95°C for 30 seconds, 62°C for 90 seconds, and 72°C for 90 seconds, and a final extension at 72°C for 10 minutes. [44]

Quantification and cleanup of PCR products

All samples were analyzed on the Agilent 2100 Bioanalyzer using the DNA 1000 kit according to the manufacturer's protocol. [47] After quantification, samples were cleaned with ExoSAP-IT™ according to the manufacturer's protocol. [48] Total concentrations of amplicons present in each sample after ExoSAP-IT™ were calculated from the Bioanalyzer data to move forward with barcoding PCR.

Library preparation

Barcoding PCR

The PCR Barcoding Kit I by ONT, which contains 12 different barcodes for multiplexing, was used. A barcoding PCR was optimized using the *TaKaRa Taq* PCR kit. The barcoding PCR was set up by combining 0.5 µl of *TaKaRa Taq* Polymerase, 5 µl of 10X PCR Buffer (with Mg^{2+}), 8 µl of a 2.5 mM each dNTP mixture, 0.5 ng of PCR product, 1 µl of barcode primers, and 30.5 µl of water. [49] The following cycling parameters were used: initial denaturation at 95°C for 3 minutes, 15 cycles of 95°C for 15 seconds, 62°C for 15 seconds, 72°C for 30 seconds, and a final extension at 72°C for 5 minutes. After PCR, amplification products were visualized and quantified on the Agilent 2100 Bioanalyzer using the DNA 1000 kit. [47] Samples were then cleaned with Agencourt® AMPure® XP beads according to the manufacturer's protocol. [50] A 1.8 volume of AMPure® beads was added to PCR products for cleanup.

End Preparation

End preparation of the barcoded library was performed according to the 1D PCR barcoding amplicons (SQK-LSK108) protocol described by ONT.[43] The library input for end prep with ONT's Ligation Sequencing Kit 1D is 1 µg of DNA in 45 µl. Because the total concentration of the samples in this study were not high enough to achieve the 1 µg input requirement, equal volumes of all samples for a sequencing run were added to the end prep reaction with a final volume of 45 µl. 5 µl of the DNA CS (control strand) from the sequencing kit, 7 µl of Ultra II End-prep reaction buffer, and 3 µl of Ultra II End-prep enzyme mix (New England Biosystems) were added to 45 µl of the pooled, barcoded library. The end-prep reaction mix was incubated at 20°C for 5 minutes and 65°C for 5 minutes.

The end-prepped libraries were then cleaned by adding 60 µl of Agencourt® AMPure® beads and incubating on a rotator mixer for 5 minutes at room temperature. Samples were then pelleted on a magnet and the supernatant was removed. Two washes with 200 µl of freshly prepared 70% ethanol were performed, removing the supernatant after 30 seconds. All ethanol was removed from the sample while on the magnet, and the beads were allowed to dry for 3 minutes. 31 µl of nuclease-free water was then added as an elution buffer, and the samples were allowed to incubate for 2 minutes at room temperature. The beads were then pelleted on the magnet, and the purified library was retained. The success of the end-prep reaction was assessed on the Agilent 2100 Bioanalyzer with the DNA 1000 kit.

Adapter ligation

Adapter ligation was performed according to the 1D PCR barcoding amplicons (SQK-LSK108) protocol.[43] The suggested input of end-prepped DNA for adapter ligation is 0.2 pmoles for a 10:1 adapter to fragment ratio. For the samples in this study, all 30 µl of end-

prepped DNA was added to the adaptor ligation reaction, regardless of concentration. 20 µl of adaptor mix (ONT) and 50 µl of Blunt/TA Ligation Master Mix (NEB) were added to 30 µl of end-prepped DNA and incubated at room temperature for 10 minutes. Excess adapter was removed by adding 180 µl of AMPure® XP beads and mixing. Samples were then incubated on a rotator mixer for 5 minutes at room temperature and placed on a magnetic rack to remove the supernatant. Two washes with 140 µl of the Adapter Bead Binding Buffer (ONT) were performed, and the supernatant was removed after 30 second incubations. The pellet was re-suspended in 15 µl of the Elution Buffer (ONT) and incubated for 10 minutes at room temperature. The beads were pelleted and the eluate was retained.

Priming and loading the flow cell

MinION flow cells (R9.4) were primed with a priming mix containing 576 µl of Running Buffer with Fuel Mix (RBF) and 624 µl of nuclease free water. 800 µl of the priming mix was added to flow cells via the priming port with the sample port closed and allowed to incubate for 5 minutes. Then 200 µl of the priming mix was added to flow cells with the sample port open. Libraries were prepared by adding 35 µl of the RBF, 2.5 µl of nuclease free water, and 25.5 µl of Library Loading Beads (LLB) to 12 µl of DNA. 75 µl of library was added to flow cells in a drop wise fashion through the sample port. The SQK-LSK108 protocol in the MinKNOW software was run along with the barcoding workflow in the EPI2ME software monitor the success of the run.

Data Analysis

Basecalling was performed on raw data files and reads were bioinformatically parsed into sample dependent files based on identification of unique barcodes. Resulting files were saved as FAST5 files with the Albacore basecalling software (ONT). FAST5 files were then converted to

FASTA files via Poretools.[51] Sequences were trimmed with Porechop, a program designed to identify ONT barcodes and adapters both at the ends of sequences and in the middle of sequences. Adapters identified at the ends of sequences are removed. When an adapter is found in the middle of the sequence, the read is considered a chimera and is split into two separate sequences. Trimming parameters were adjusted to better fit short reads. Adapter sequences were forced to have at least 98% identity to at least 10,000 sequences in the file in order to be counted as present (--adapter_threshold 98). The program checked for adapters in the 150 bases at the ends of each sequence (default setting) and removed sequences that had 85% similarity to it (--end_threshold 85), without trimming any extra bases (--extra_end_trim 0). Adapters in the middle of the sequence had to have 98% similarity to the adapter sequences to be split (--middle_threshold 98). Any sequences after the split that were less than 100 bp in length were removed (--min_split_read_size 100). No additional bases were removed on either side of the adapters (--extra_middle_trim_good_side 0 and --extra_middle_trim_bad_side 0). All other parameters were kept at the default settings.

FASTA files were then uploaded into the CLC Genomics Workbench Software. All sequences less than 150 bp were filtered out with the trimming tool to prevent analysis of primer dimers. Remaining sequences were then mapped to the genes included in the multiplex. All reference gene sequences were obtained from GenBank.

CHAPTER THREE: RESULTS

DNA/RNA co-extraction

The DNA/RNA co-extraction method appeared to be sufficient to obtain full DNA profiles for most of the undiluted body fluid samples, as well as the mixture samples and some of the diluted body fluid samples. Full STR profiles were obtained for all semen and menstrual blood samples from Lee Biosolutions™ (Table 3). Full STR profiles were obtained for six of the eight saliva samples, with the remaining two having 50% and 78.6% of complete profiles. Full profiles were obtained for seven of the eight blood samples and seven of the eight vaginal fluid samples, with the remaining two samples having greater than 90% of a complete profile.

Table 3. Average DNA quantities and percentage profile completeness based on autosomal STR loci for semen, saliva, blood, vaginal fluid, and menstrual blood samples.

Body Fluid	N	Average DNA Quantity (ng)	Profile Completeness (%)
Semen	8	0.380 ± 0.260	100.0 ± 0.000
Saliva	8	0.083 ± 0.120	91.07 ± 18.21
Blood	8	0.028 ± 0.020	99.12 ± 02.53
Vaginal Fluid	8	03.90 ± 02.93	98.81 ± 03.37
Menstrual Blood	8	0.270 ± 0.160	100.0 ± 0.000

Complete or nearly complete profiles (greater than 90% of a profile) were obtained for 77.8% of the 1:10 dilutions of body fluids (Table 4). For semen, full profiles were obtained for all three of the 1:10 dilution replicates. Complete profiles were obtained for only two of the three 1:10 saliva dilution replicates. No complete profiles were obtained for the 1:10 blood dilutions, which had an average of 88.9% (± 5.0%) of a complete profile. No full profiles were obtained for 1:100 dilutions. For semen, one of the 1:100 dilutions had 92.9% of a complete

profile, with the remaining two having less than 30% of a complete profile. For saliva samples, one of the dilutions had no alleles present, while the other two had 33% and 17 % of a complete profile. For blood, one of the 1:100 dilution samples had 95.2% of a complete profile. The other samples had less than 20% of a complete profile. For 1:1,000 semen dilutions, one sample had no alleles present, while the other two had less than 20% of a profile. No profiles were obtained for any of the 1:1,000 dilutions of saliva. For the 1:1,000 dilutions of blood, two samples had no alleles present, while the other had less than 5% of a complete profile.

Table 4. Average DNA sample quantities and percentage profile completeness based on autosomal STR loci for 1:10, 1:100, and 1:1,000 dilutions of semen, saliva, and blood.

	Body fluid	N	Dilution factor	Average DNA quantity (ng)	Profile Completeness (%)
1	Semen	3	1:10	0.040000 ± 0.036000	100.0 ± 0.000
2		3	1:100	0.000650 ± 0.000820	46.83 ± 39.94
3		3	1:1,000	0.000033 ± 0.000058	07.14 ± 08.58
4	Saliva	3	1:10	0.007000 ± 0.005200	75.40 ± 42.61
5		3	1:100	0.000200 ± 0.000200	16.67 ± 16.67
6		3	1:1,000	–	–
7	Blood	3	1:10	0.002000 ± 0.001300	88.89 ± 04.96
8		3	1:100	0.000520 ± 0.000560	39.68 ± 48.64
9		3	1:1,000	–	00.79 ± 01.37

For the mixture samples, full profiles were obtained for the major contributors (vaginal swab donors) for all samples (Table 5). An average of 57.9% (± 11.3%) of a complete profile was obtained for the minor contributors (semen donors) from vaginal swabs with 7.5 µl of semen. An average of 96.03% (± 5.0%) of a complete profile was obtained for minor contributors from swabs with 15 µl of semen. Complete profiles were obtained for minor contributors from two of the three swabs with 30 µl of semen, with the remaining replicate having 95.2% of a complete profile.

Table 5. Average DNA quantities and percentage profile completeness based on autosomal STR loci for vaginal swabs with 7.5 µl, 15 µl, and 30 µl of semen.

Semen volume (μl)	N	Average DNA quantity (ng)	Contributor	Profile Completeness (%)
7.5	3	08.83 ± 05.16	Major Minor	100.0 ± 0.000 57.94 ± 11.25
15	3	13.29 ± 06.63	Major Minor	100.0 ± 0.000 96.03 ± 04.96
30	3	10.05 ± 01.42	Major Minor	100.0 ± 0.000 98.41 ± 02.75

After extraction, analysis of RNA samples on the Bioanalyzer with the RNA 6000 Pico Kit revealed all RNA samples to be highly degraded. The average RIN number for 8 semen, saliva, and blood samples and ten menstrual blood samples was 1.93. 12 of the 34 samples analyzed with the RNA 6000 Pico Kit could not be assigned an RIN number. Despite the apparent degradation, all samples were taken forward for sequencing. Because amplification was seen in RNA samples after PCR, it appears that the RNA quantification method is not informative of the quality and quantity of the RNA in this workflow. For the rest of the samples in this study, the RNA quantitation step was removed, and samples were reverse transcribed directly after extraction and DNase treatment.

mRNA analysis

Marker specificity for short targets

The performance of primers for all mRNA markers was assessed in singleplex to ensure no cross-reactivity of body fluids. The success of PCR was determined on the Agilent 2100 Bioanalyzer. The lengths of amplicons produced for each sample were compared to expected amplicon lengths. In singleplex reactions targeting short amplicons, no cross reactivity of body fluid specific markers was observed in any of the samples according to the amplicon length data.

Amplicons of the proper length were observed for all markers in their respective target body fluid, except for MMP7 and MMP11. Amplicons of the proper length for G6PDH were obtained in the multiplex reaction, but not the singleplex reaction. The MMP7 and MMP11 markers were not amplified in the menstrual blood samples obtained from Lee Biosolutions™. However, an amplicon matching the expected length of MMP7 was present in PCR products obtained from a vaginal swab that may have had menstrual blood present. The 18S rRNA marker has shown amplification in the multiplex PCR products for reagent blanks and non-template controls, but not in any of the singleplex reactions. However, the signal is low in the reagent blanks in comparison to the signal in body fluid samples. Typical Bioanalyzer results after the multiplex PCR for semen, saliva, blood, and vaginal fluid samples are shown in Figure 2.

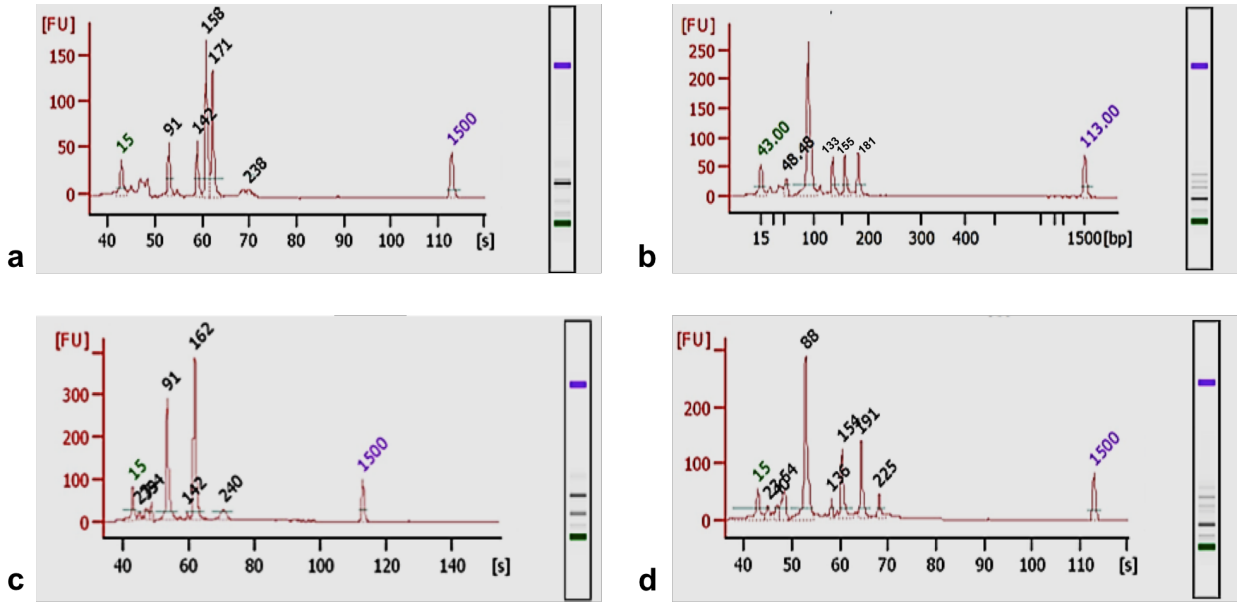

Figure 2. Bioanalyzer results for (a) a semen sample showing amplicons present at the approximate expected lengths for PRM1 (142 bp), SEMG1 (171 bp), 18S rRNA (158 bp), and G6PDH (238 bp), (b) a saliva sample showing amplicons present at the approximate expected lengths for STATH (133 bp), HTN3 (181 bp), and 18S rRNA (155 bp), (c) a blood sample showing amplicons present at the approximate expected lengths for ALAS2 (142 bp), a potential combined peak for HBA and18S rRNA (162 bp), and G6PDH (240 bp), and (d) a vaginal fluid sample showing amplicons present at the approximate expected lengths for MYOZ1 (136 bp), CYP2B7P1 (191 bp),18S rRNA (154bp), and G6PDH (225 bp).

For semen samples, amplicons matching the approximate length for PRM1 and SEMG1 were present in five of the eight samples, and amplicons matching the approximate length for PRM1 were present in the other three samples (Table 6). Amplicons matching the lengths of both HTN3 and STATH were seen in seven of the saliva samples, and an amplicon matching only HTN3 was seen in one sample. Amplicons matching ALAS2 were seen in all of the blood samples. Due to the closeness in length of the HBA and18S rRNA markers, the presence of the HBA marker cannot be inferred from fragment lengths in the multiplex. Fragments close to the length of the vaginal fluid markers were seen in most of the vaginal fluid samples. There was no amplification of the menstrual blood markers seen in any of menstrual blood samples.

Table 6. Presence of gene markers according to length data in eight semen, blood, saliva, and vaginal fluid samples and in ten menstrual blood samples (eight liquid samples and two vaginal swabs).

		PRM1	SEMG1	HTN3	STATH	HBA	ALAS2	CYP2B7P1	MYOZ1	MMP7	MMP10
1	Semen	8/8	5/8	0/8	0/8	0/8	0/8	0/8	0/8	0/8	0/8
2	Saliva	0/8	0/8	8/8	7/8	0/8	0/8	0/8	0/8	0/8	0/8
3	Blood	0/8	0/8	0/8	0/8	8/8	8/8	0/8	0/8	0/8	0/8
4	Vaginal fluid	0/8	0/8	0/8	0/8	0/8	0/8	7/8	6/8	0/8	0/8
5	Menstrual blood	0/10	0/10	0/10	0/10	0/10	0/10	1/10	0/10	0/10	0/10

According to BLAST results, Sanger sequencing performed on one sample per body fluid type, as well as one of each of the dilution series samples and mixture samples, showed the presence of both semen markers in the semen sample. HTN3 was present in the saliva sample, but no sequence was obtained for STATH. HBA was present in the blood sample, but no sequence was obtained for ALAS2. Sequences matching both CYP2B7P1 and MYOZ1 were obtained from the vaginal fluid sample. No sequence data was obtained for either of the menstrual blood gene markers.

Sensitivity of short targets

For the semen dilution series, amplicons matching the length of PRM1 were observed in all three of the 1:10 and 1:100 dilution samples, and was observed in one of the 1:1,000 dilution samples (Table 7). Amplicons near the expected length of SEMG1 were observed in one of the three 1:10 and 1:1,000 dilution samples and in all three of the 1:100 dilutions. For the saliva dilution series, amplicons matching the length of HTN3 were seen in all three of the 1:10 dilution samples, but in none of the 1:100 or 1:1,000 dilutions. An amplicon near the expected length of STATH was seen in only one of the 1:10 dilution samples, with no amplification of STATH apparent in the 1:100 or 1:1,000 samples. For the blood dilution series, an amplicon matching the length of HBA was seen in only one of the 1:10 dilution samples. There was no amplification of HBA in the 1:100 or 1:1,000 dilutions. Additionally, there was no apparent amplification of ALAS2 in any of the samples. For the semen/vaginal fluid mixture series, amplicons close to the expected lengths of PRM1 and SEMG1 were seen in all samples. Amplicons close to the expected length of CYP2B7P1 were seen in all samples, but MYOZ1 did not appear to amplify in any of the samples.

Table 7. Presence of gene markers according to length data in (a) a tenfold dilution series of semen, saliva, and blood, and in (b) a mixture series with vaginal swabs combined with decreasing volumes of semen (30 µl, 15 µl, and 7.5 µl).

a

		PRM1	SEMG1	HTN3	STATH	HBA	ALAS2
Semen	1:10	3/3	1/3	0/3	0/3	0/3	0/3
	1:100	3/3	3/3	0/3	0/3	0/3	0/3
	1:1,000	1/3	1/3	0/3	0/3	0/3	0/3
Saliva	1:10	0/3	0/3	3/3	1/3	0/3	0/3
	1:100	0/3	0/3	0/3	0/3	0/3	0/3
	1:1,000	0/3	0/3	0/3	0/3	0/3	0/3
Blood	1:10	0/3	0/3	0/3	0/3	1/3	0/3
	1:100	0/3	0/3	0/3	0/3	0/3	0/3
	1:1,000	0/3	0/3	0/3	0/3	0/3	0/3

b

		PRM1	SEMG1	CYP2B7P1	MYOZ1
1	30 μl Semen	3/3	3/3	3/3	0/3
2	15 μl Semen	3/3	3/3	3/3	0/3
3	7.5 μl Semen	3/3	3/3	3/3	0/3

According to the BLAST results, Sanger sequencing performed on one of the 1:10 semen dilution samples showed the presence of both semen markers. For the 1:10 saliva dilution sample, a sequence matching HTN3 was observed, but no sequence was obtained for STATH. A sequence matching HBA was seen in the 1:10 blood dilution sample analyzed, but no sequence was obtained for ALAS2. For the mixture sample (vaginal swab with 30 µl of semen), sequences for both of the semen markers and for both of the vaginal fluid markers were obtained.

Marker specificity for long targets

Primers targeting longer fragments were tested in singleplex reactions to ensure no occurrence of cross reactivity. There was no non-specific amplification seen in off-target body fluids. However, within target body fluids, there was amplification of fragments approximately twice as long as expected fragments. This appeared to occur more often in the semen samples

than other samples. This non-specific amplification within body fluid targets also occurred in the multiplex. Because this appeared to occur only within target body fluids, the primers in the multiplex were considered to be body fluid specific and were used to amplify all samples prior to sequencing.

All FAST5 files from the passed and skipped reads folders generated by the MinION were basecalled with Albacore, trimmed with Porechop, and loaded into the CLC Genomics Workbench software to filter out reads less than 150 bp in length and to map reads to the reference genes. There was a lot a variability in total read counts for each sample in a run. Due to samples not being normalized prior to sequencing preparation, some samples had many more reads than others. Samples with a less than 2,000 total reads were removed prior to analysis. A threshold of 2,000 was chosen based on thresholds set by Hanson et al., in which they count any gene marker with at least 500 reads as present. [52] Each sample should contain reads for four different genes (two body fluid specific and two housekeeping), so the threshold was set at 2,000 reads. Because the number of reads from sample to sample and from run to run is highly variable, the relative number of reads of each gene in each body fluid sample was calculated by dividing the number of reads of the gene by the total number of reads for the sample. For each body fluid type, the average relative read counts of each gene across all samples was calculated. The relative read counts of body fluid specific genes are represented as percentages of total reads. For semen samples, PRM2 appeared to be the most successful marker, contributing to an average of 87% of total reads (Figure 3). PRM2 was the most consistently and highly expressed marker in semen samples. The expression of SEMG1 was less consistent. SEMG1 was the second most highly expressed gene in semen samples, contributing to an average of 11% of the total number of reads, but expression was not consistent among all semen samples. In two

semen samples, SEMG1 comprised less than 0.5% of total reads. Expression of off-target genes was minimal. The number of reads of off-target genes in all eight samples was less than 0.5% of the total number of reads. Like SEMG1, the housekeeping genes, B2M and UBC, were not expressed uniformly across all samples. Only one sample had a significant number of reads of B2M (2,381 reads), while all other samples had less than 50. Two samples had a significant number of UBC reads (2,670 and 2,994 reads), but all other samples had less than 10.

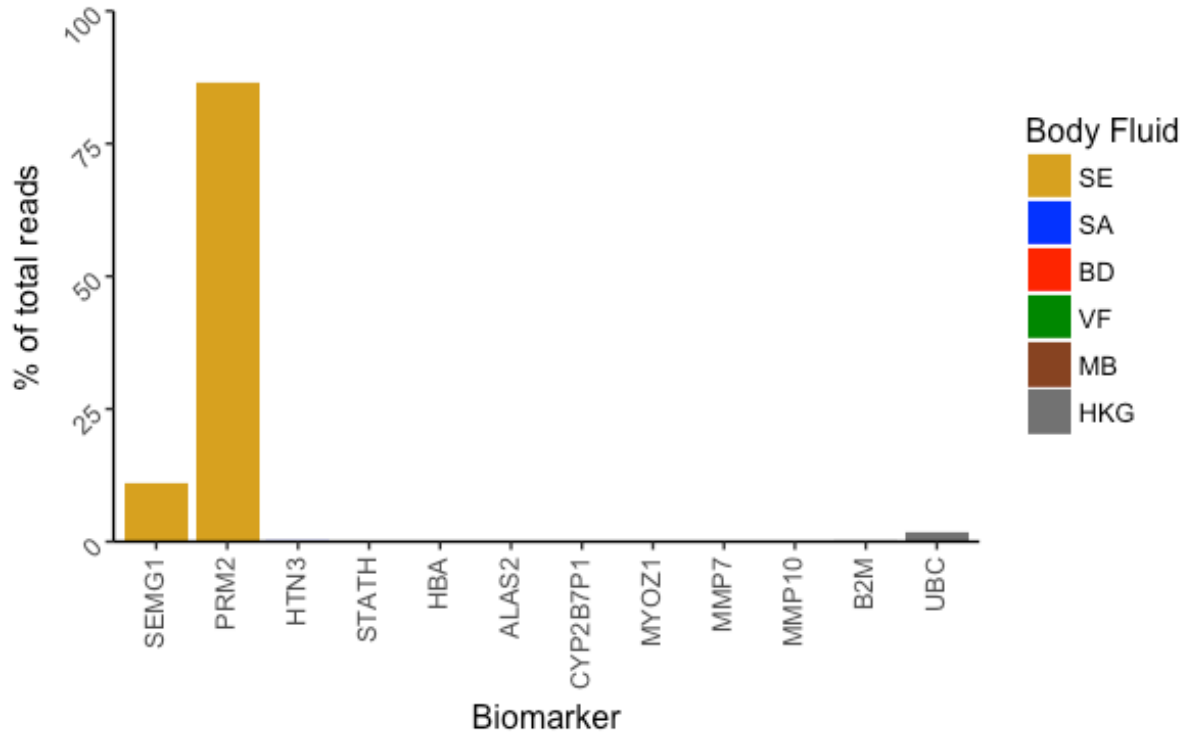

Figure 3. Average percentages of relative reads for all targeted genes in semen samples (n=8). The yellow bars represent expression of the semen specific genes and the gray bar represent housekeeping gene expression.

Two saliva samples had less than 2,000 total reads and were eliminated from analysis. For the remaining saliva samples, HTN3 appeared to be the most highly and uniformly expressed gene in the multiplex, contributing to an average of 62% of all reads (Figure 4). However, one

sample had only 3 reads of HTN3. STATH, which contributed to 17% of all reads, was not expressed as uniformly. Three samples had 0 reads of STATH. Of the remaining three samples, one had 90,927 reads of STATH, and the other two had 1,526 and 3,514 reads. Like in the semen samples, the housekeeping genes were not expressed uniformly across all saliva samples. For three of the samples, the number of reads of B2M was less than 30, and four of the samples had less than 10 reads of UBC. One sample had 24,195 reads of B2M, which contributed to the majority of the total reads across all samples. Only one sample had a significant number of reads of UBC (1,728 reads).

The number of off-target reads was slightly higher than in the semen samples. There were a greater number of PRM1 and SEMG1 reads for several samples. However, this was not seen uniformly across all saliva samples. Increased number of off-target reads appears to be correlated with reuse of the flow cells. Samples that had a significant number of reads of semen genes (greater than 0.5% of total reads) were run on a washed flow cell that had sequenced semen samples prior to loading the saliva samples, so the unexpected number of reads is most likely due to library carryover from a previous run. The number of reads of other off-target genes (blood, vaginal fluid, and menstrual blood) were less than 0.5% of the total.

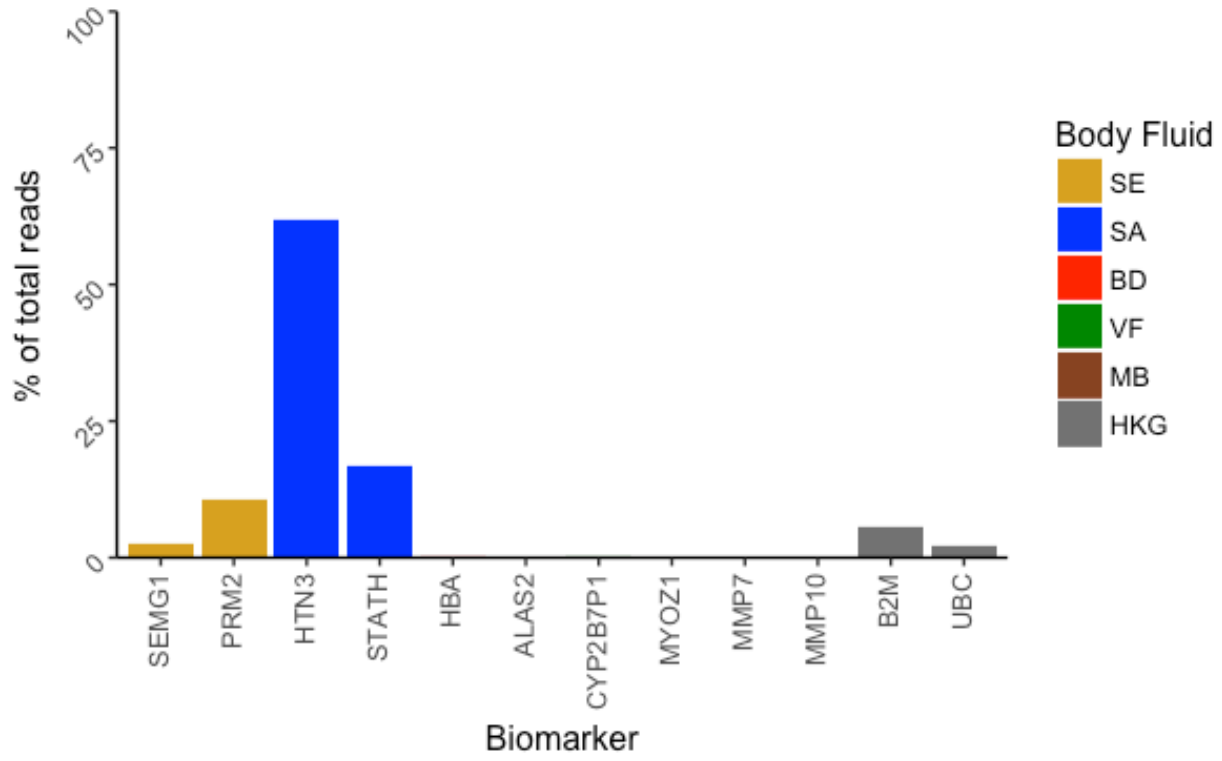

Figure 4. Average percentages of relative reads for all targeted genes in saliva samples (*n*=*6*). The blue bars represent expression saliva specific genes, while the gray bars represent housekeeping gene expression.

In the blood samples, HBA was the most highly expressed, contributing to 59% of all reads (Figure 5). All samples had at least 1,000X coverage of HBA. ALAS2 was not expressed as highly as HBA, contributing to only 10% of all reads. In one of the samples, ALAS2 comprised less than 0.5% of the total number of reads. The housekeeping genes were more uniformly expressed in blood samples than in semen and saliva samples. B2M comprised more than 0.5% of total reads in all samples, with all but one sample having 1,000X coverage. UBC was not expressed very highly in blood samples. Six of the eight samples had less than 500 reads of UBC. There was also unexpectedly high expression of semen and saliva genes in some of the samples. This appears to be due to residual library left after washing the flow cell of a previous sequencing run, as the pattern was not seen uniformly across all samples.

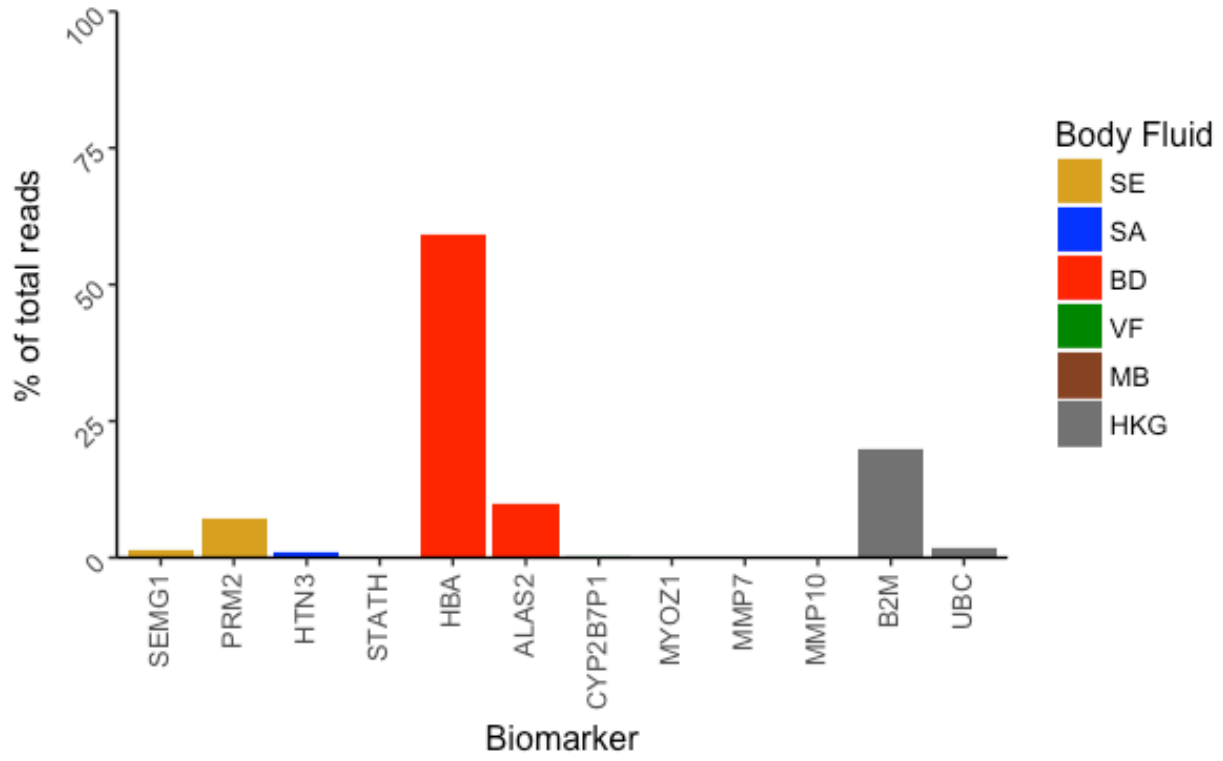

Figure 5. Average percentages of relative reads for all targeted genes in blood samples (*n*=*8*). The red bars represent expression of blood specific genes, while the gray bars represent housekeeping gene expression.

One of the vaginal fluid samples had less than 2,000 total reads and was removed from the analysis. For the remaining seven samples, CYP2B7P1 appeared to be highly and uniformly expressed in all samples, contributing to 34% of all reads (Figure 6). However, in one sample, the number of CYP2B7P1 reads was only 124. In the other six samples, there was more than 10,000X coverage of CYP2B7P1. MYOZ1 was not observed in vaginal fluid samples. Highest number of reads obtained for MYOZ1 was 958. MYOZ1 comprised less than 0.5% of total reads in all other samples. Both housekeeping genes were heavily expressed in vaginal fluid samples. All samples had greater than 5,000X coverage of UBC, contributing to 51% of all reads. B2M was more sporadically expressed than UBC, with three of the samples having less than 500 reads of B2M. There were an unexpected number of reads of SEMG1, HBA, and MMP7 in one

vaginal fluid sample. In all other samples, the number of off-target reads was less than 0.5% of the total.

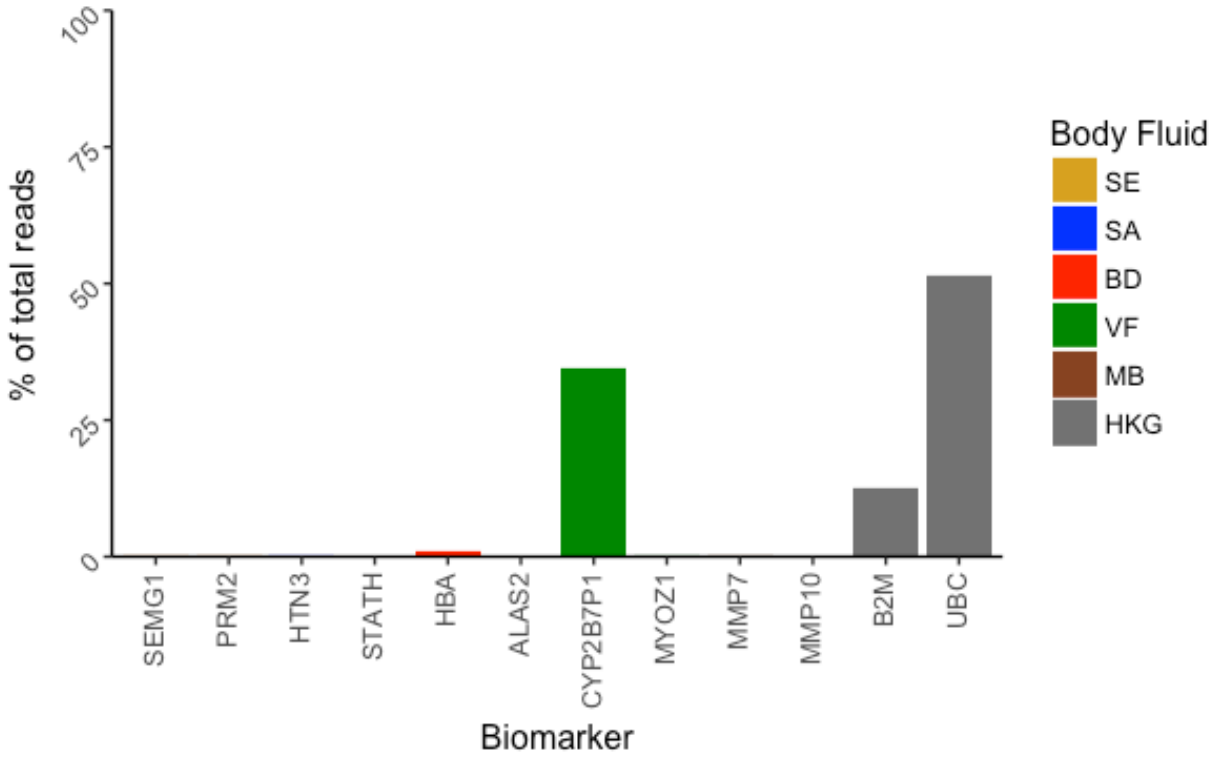

Figure 6. Average percentages of relative reads for all targeted genes in vaginal fluid samples (*n*=7). The green bars represent expression of vaginal fluid specific genes, and the gray bars represent housekeeping genes expression.

In the menstrual blood sample analyzed, the two most highly expressed genes were HBA and B2M, which comprised 34.9% and 30.8% of all reads, respectively (Figure 7). UBC comprised 17% of all reads and the vaginal fluid gene CYP2B7P1 comprised 8.4% of all

reads. A significant number of reads was obtained for the menstrual blood marker MMP7 (2,925 reads), but no reads of MMP10 were observed.

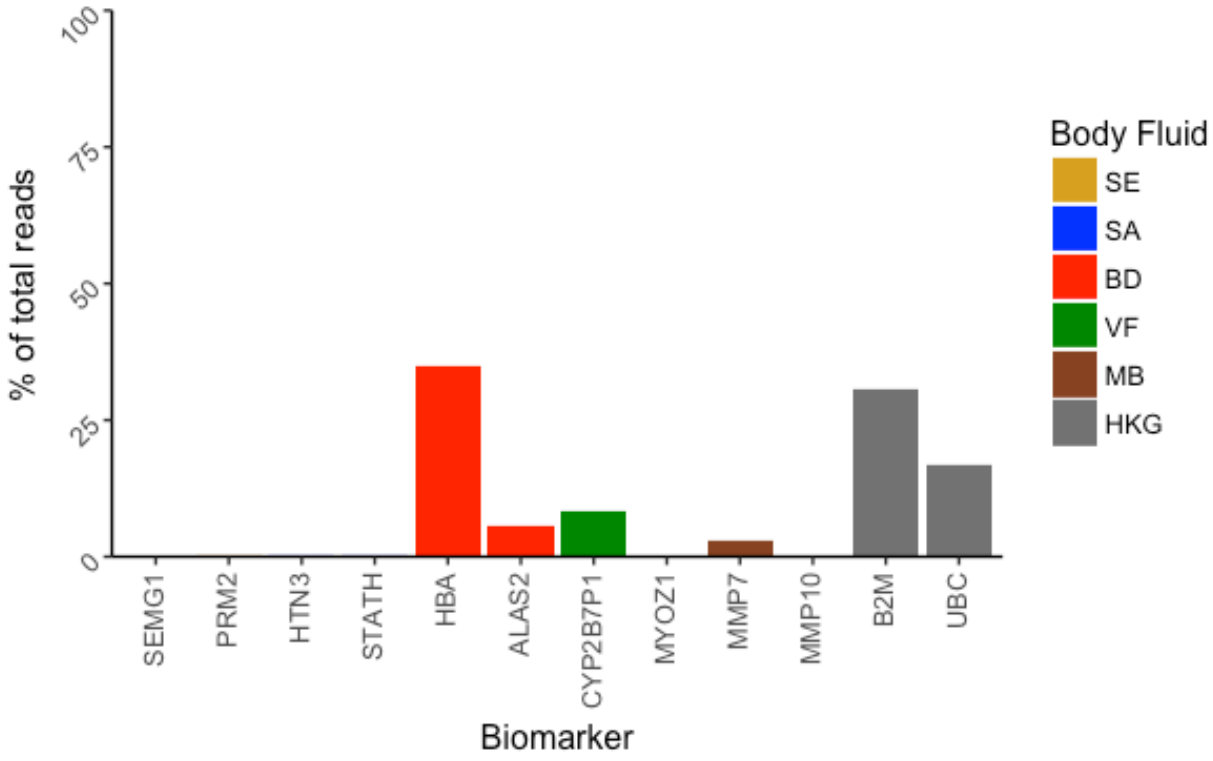

Figure 7. Percentages of relative reads for all targeted genes in one menstrual blood sample. The brown bars represent expression of menstrual blood genes, green bars represent expression of vaginal fluid specific genes, red bars represent expression of blood genes, and gray bars represent expression of housekeeping genes.

The number of on and off-target reads for all samples, as well as the percentage of off-target reads were compared (Table 8). For all body fluid types, the numbers of on-target reads (body fluid specific and housekeeping gene reads) comprised the majority of all reads. Semen and menstrual blood samples had the lowest percentage of off-target reads. There was unexpected expression of off-target genes in saliva, blood, and vaginal fluid samples. However, both saliva and blood samples were run on flow cells that had been previously used and washed.

There appears to be residual library present after washing of the flow cells, which can account for the increased number of off-target reads. For the vaginal fluid samples, one sample comprised the majority of all off-target reads. An unexpected number of HBA, MMP7, and PRM2 gene reads were seen in one sample, but the pattern was not seen uniformly across all seven samples.

Table 8. Total number of reads, as well as the total number of on and off-target reads and the percentage of off-target reads, across all body fluid types.

	N	Total reads	On target reads	Off target reads	% of off target reads
Semen	8	269,534	269,229	305	0.11
Saliva	6	301,362	297,355	4,007	1.33
Blood	8	178,883	156,926	21,957	12.27
Vaginal fluid	7	880,421	854,914	25,507	2.90
Menstrual blood	1	100,803	100,543	260	0.26

Average relative read counts of body fluid specific genes for each body fluid type are represented as percentages (Table 9). For semen samples, semen specific genes contributed to 97.7% of reads, with little contribution from any other genes. For saliva samples, the saliva specific genes contributed most of the reads, with more than expected contribution from semen genes due to flow cell contamination. For blood, the largest contribution of reads was from blood specific markers, with more than expected contribution from semen genes, which was also due to flow cell contamination. In both vaginal fluid and menstrual blood samples, the largest percentage of contribution was from housekeeping genes. The second largest contribution was from vaginal fluid genes in these samples. In menstrual blood, the contribution from menstrual blood genes was low. However, contribution from blood and housekeeping genes was high.

Table 9. Percent contribution of body fluid specific gene reads based on average relative read counts. Cells highlighted in yellow represent the percent contribution of expected genes (biomarkers) for each body fluid type, and cells highlighted in blue represents the percent contribution of housekeeping genes.

Biomarkers	Semen $N = 8$	Saliva $N = 6$	Blood $N = 8$	Vaginal Fluid $N = 7$	Menstrual Blood $N = 1$
Semen	97.7	13	8.3	0.2	0.1
Saliva	0.1	78.9	0.9	0.3	0.2
Blood	0	0.3	69	1	40.7
Vaginal Fluid	0	0.1	0.1	34.6	8.4
Menstrual Blood	0	0	0	0.1	2.9
Housekeeping	2.1	7.8	21.8	63.9	47.7

Sensitivity of long targets

The 1:10, 1:100, and 1:1,000 dilutions of semen, blood, and saliva were sequenced in triplicate and numbers of reads for each gene were determined (Table 10). In the semen dilution series, PRM2 appeared to be the most successful of the semen markers. In 1:10 dilutions, PRM2 was present in all samples and comprised the majority of all reads, while there were less than 50 reads of SEMG1 across all samples. Reads of semen genes in the 1:100 dilutions were also observed. More reads of SEMG1 were seen in the 1:100 dilution samples than in the 1:10 dilutions, with 15,283 reads in one replicate and 4,851 reads in another. Over 4,000 reads of PRM2 were also observed in two of the 1:100 dilution replicates. The read counts of all other genes, including housekeeping genes, were less than 500. Low read counts were obtained for all genes in the 1:1,000 replicates, which had less than 20 reads of semen genes.

This assay appears to be sensitive enough to obtain saliva gene reads in 1:10 dilutions of saliva. HTN3 was present in the 1:10 dilutions of saliva. Two of the samples had more than 1,000X coverage, with little to no contribution from other genes, including housekeeping genes. No reads of STATH were observed in any of the samples. Very low read counts were obtained

for 1:100 and 1:1,000 dilutions of saliva. One of the 1:100 dilution replicates had higher than expected reads of CYP2B7P1 and UBC, which appears to be due to contamination of the flow cell with residual library that contained vaginal fluid samples.

Reads of blood specific genes (particularly HBA) were able to be obtained down to 1:1,000 dilutions of blood. Over 1,000X coverage of HBA was seen in all blood dilution samples except for one of the 1:1,000 dilution replicates. Little to no reads of the other blood specific gene, ALAS2, were observed. There were an unexpectedly high number of reads of semen genes in some of these samples, as well as reads of the vaginal fluid gene CYP2B7P1 in some samples. This appears to be correlated with contamination from residual library on the flow cell. In some samples, there were a large number of reads of the housekeeping genes. However, samples with more housekeeping gene reads had the same barcodes as vaginal fluid samples run previously on the flow cell, and because the high expression of housekeeping genes in vaginal fluid samples, it is likely that this is contamination.

Table 10. Total number of reads for each sample in the dilution series. Numbers highlighted in green comprise >10% of total reads for the sample, while numbers highlighted in yellow represent <5% of total reads and numbers in white comprise <1% of all reads.

Body Fluid	Dilution Factor	SEMG1	PRM2	HTN3	STATH	HBA	ALAS2	CYP2B7P1	MYOZ1	MMP7	MMP10	B2M	UBC	Total
Semen	1:10	33	7,556	0	0	14	0	104	2	2	0	123	286	8,120
		14	3,151	0	0	10	0	32	1	0	0	47	81	33,36
		0	35,383	1	0	0	0	0	0	0	0	0	0	35,384
	1:100	15,283	23	0	0	7	0	70	1	1	0	116	152	15,653
		4,851	10,716	0	0	20	0	89	0	1	0	121	189	15,987
		0	4,666	1	0	0	0	0	0	0	0	0	0	4,667
	1:1,000	12	7	0	0	8	0	27	0	2	0	63	69	188
		0	10	0	0	0	0	0	0	0	0	0	0	10
		0	19	1	0	0	0	0	0	0	0	0	0	20
Saliva	1:10	0	7	1,293	0	0	0	0	0	0	0	0	0	1,300
		0	9	1	0	0	0	0	0	0	0	0	0	10
		0	8	1,933	0	0	0	0	0	0	0	0	0	1,941
	1:100	0	8	1	0	0	0	0	0	0	0	0	0	9
		0	9	0	0	0	0	0	0	0	0	0	0	9
		4	6	0	0	19	0	854	1	0	0	26	2,212	3,122
	1:1,000	0	10	3	0	0	0	0	0	0	0	0	0	13
		0	1	2	0	0	0	0	0	0	0	0	0	3
		14	8	0	0	21	0	21	0	0	0	29	69	162
Blood	1:10	1	2	0	0	3,099	0	3,649	71	0	0	706	2,540	10,068
		376	14	0	0	18,535	0	13,494	0	323	0	31,168	18,254	82,164
		7	4	0	0	6,314	146	4	0	0	0	11	27	6513
	1:100	10	8	0	0	10,546	0	21	0	0	0	54	18,466	29,105
		17	2,480	0	0	2,624	0	24	0	1	0	27	49	5,222
		6	1,001	0	0	6,998	0	5	0	0	0	12	18	8,040
	1:1,000	6	11	0	0	29	0	2,478	0	2	0	1,466	10,337	14,329
		4,571	9	0	0	2,426	0	11	0	1	0	24	33	7,075
		1,333	3,406	0	0	4,314	0	17	0	0	0	31	45	9,146

Both semen and vaginal fluid genes, as well as housekeeping genes, were observed in the vaginal fluid/semen mixtures. More than 1,000X coverage of SEMG1 was obtained for one of the replicates that had 7.5 µl and 15 µl of semen, and for two of the replicates that had 30 µl of semen. Decreasing the volume of semen on swabs appeared to lead to overall decrease in semen gene reads, with only one of the samples with 7.5 µl of semen having a significant amount of reads. More than 1,000X coverage of PRM2 was seen in one of the samples with 7.5 µl of semen, and in two of the samples with 15 µl and 30 µl of semen. More than 1,000X coverage of the vaginal fluid marker CYP2B7P1 was observed for five of the nine samples. Little to no reads of the other vaginal fluid marker, MYOZ1, were obtained. Housekeeping genes were heavily expressed in nearly all samples. The total number of reads obtained during this run was lower than previous runs. These samples were run on a flow cell that had previously completed two sequencing runs.

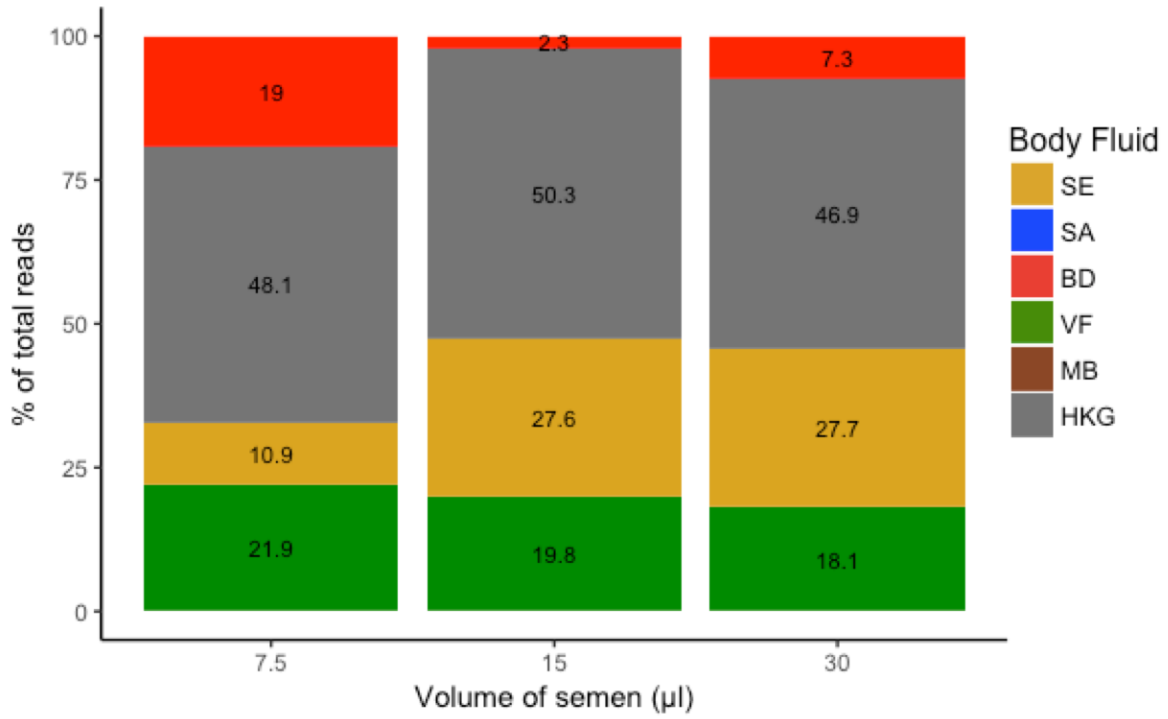

Figure 8. Average percentages of relative reads for all targeted genes in vaginal fluid swabs with 7.5 µl (n=3), 15 µl (n=3), and 30 µl (n=3) of semen.

MinION performance

Despite the 1 µg of recommended DNA input for the 1D Ligation Sequencing Kit, the MinION appears to be able to generate high quality data and over 1,000X coverage of target genes with little to no off-target reads for most samples. Contamination of flow cells with residual library after washing is apparent, which led to a greater number of off-target reads for some samples. Flow cells were able to be washed and reused 2-4 times, although some flow cells had less than the recommended 800 single active pores available after only one use and were therefore unusable. In order to assess the loss of pore function with use of the flow cell, the number of reads passing through each pore on one flow cell throughout each consecutive sequencing run was obtained with Poretools (Figure 9).[51] There was a clear gradual loss in the occupancy of pores with each consecutive run. The first run on the flow cell (Figure 9 a) had a

much higher occupancy than the second (Figure 9 b) and third (Figure 9 c) runs. Although ONT states that flow cells can be washed and reused 3-4 times, the sequencing yield appears to be negatively affected by reuse. The total number of reads obtained in the first run on the flow cell was 257,038. Most of the samples in the first run were vaginal swab samples, which had a high number of reads overall. The total number of reads obtained in the third run on the same flow cell was 122,474. Vaginal swabs samples also comprised a majority of the samples in the third run. Despite having similar sample types, the first run on the flow cell gathered over 2 times as many reads as the third run.

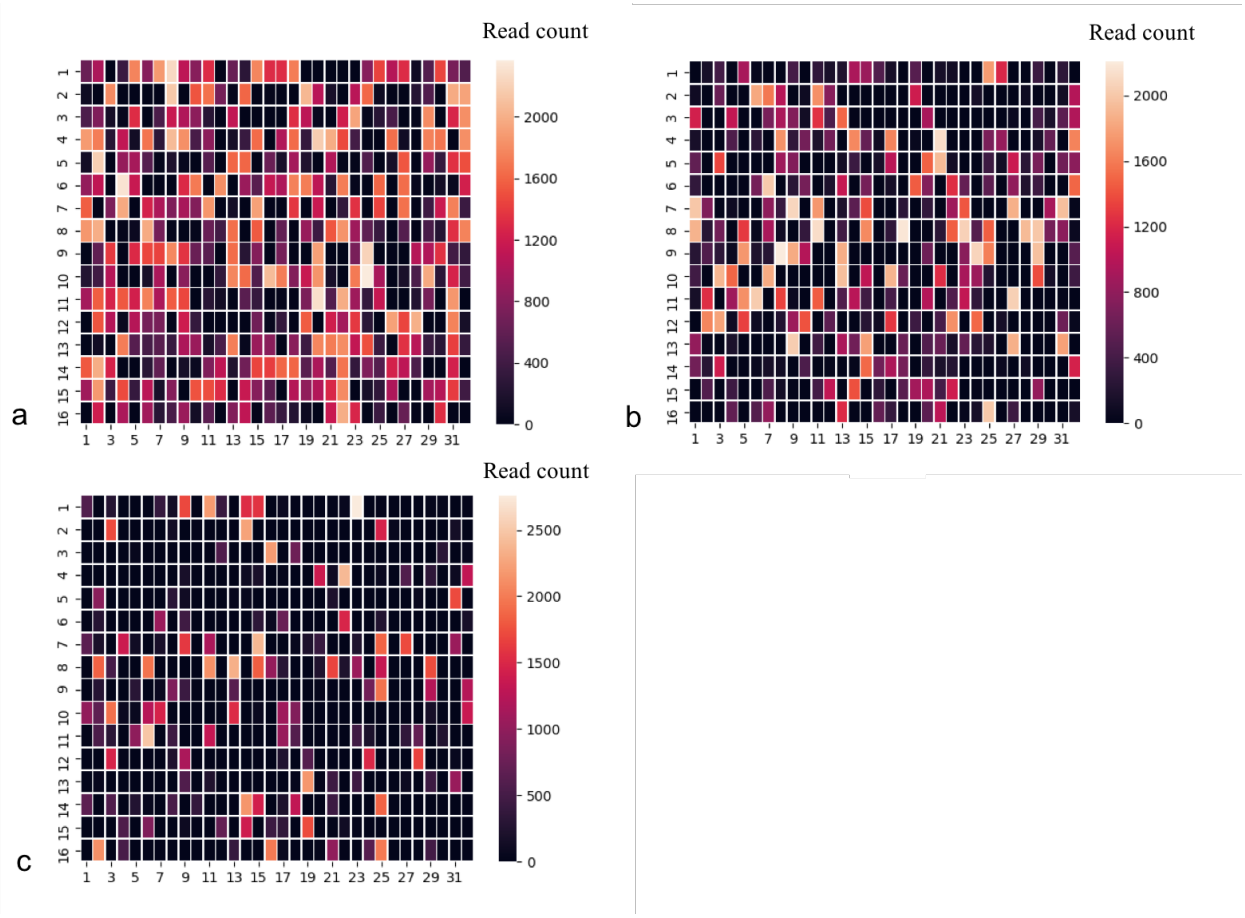

Figure 9. Heat map showing the number of reads passed through each nanopore on a flow cell during (a) the first run, (b) the second run, and (c) the third run. Blocks with lighter colors

represent nanopores that obtained a greater number of reads, while blocks with darker colors represent nanopores that obtained fewer reads.

CHAPTER FOUR: DISCUSSION

DNA/RNA co-extraction

According to the results of DNA quantification, the DNA/RNA co-extraction method using a combination of the RNeasy® Kit from Qiagen® and the PrepFiler® Kit from Life Technologies™ appears to be sufficient to obtain enough DNA to generate full STR profiles. Additionally, partial profiles and some complete profiles were obtained for the samples in the dilution series. For the mixture samples, full profiles for both the major and minor contributors were able to be obtained. However, whether or not there is decreased DNA yield with this method in comparison to the yield obtained with the Prepfiler® Kit alone is unknown. In order to validate this method for use in crime labs, the DNA yield obtained with the adapted Prepfiler® protocol used in this study should be compared to the yield obtained from extraction with the Prepfiler® Kit according to the manufacturer's protocol. Assessment of the performance of some commercial DNA/RNA co-extraction methods have shown that there is currently no kit that offers optimal extraction of both DNA and RNA. [7, 8] If adding 350 µl RLT Buffer from the RNeasy Kit to swabs for the 1 hour incubation at 56°C instead of the Prepfiler® Buffer does not decrease DNA yield, it could be an efficient co-extraction method for use in crime labs. The RNeasy® Kit has the ability to be automated, which would be time-efficient for analysts and would decrease the potential for contamination. With automation, the complete purification of DNA and RNA from extract fractions could be done simultaneously.

RNA quantification with the Agilent RNA 6000 Pico Kit did not appear to be indicative of the true quality of the RNA samples in this study. All samples were shown to be highly degraded, and RIN values were not able to be obtained for many of the samples. Haas et al.

(2009) have suggested that both the Nanodrop and Bioanalyzer RNA kits are unreliable measures of quality and quantity of RNA in forensic samples, having achieved positive results for mRNA markers despite RIN numbers showing highly or completely degraded RNA.[53] In a collaborative EUROFORGEN/EDNAP study, authors suggest using the Quant-iT RiboGreen RNA kits with the fluorescence microplate reader (Thermo Fisher Scientific), QuantiFluor RNA System (Promega), or the Quant-iT RNA assay with the Qubit (Thermo Fisher Scientific) instead, as the Bioanalyzer generally produces low RIN numbers for forensic type samples.[54] For the purposes of this workflow, RNA quantification may not be necessary, as the main reason for quantification of RNA is for the cDNA synthesis reaction. The RNA input for the Protoscript® First Strand cDNA Synthesis Kit is up to 1 µg of RNA, which would likely never be achieved with forensic RNA samples.[37] However, for sequencing kits that require RNA rather than cDNA as the input for sequencing, other quantification kits should be explored.

mRNA analysis

Marker specificity

Bioanalyzer results for singleplex PCR reactions targeting both short and long targets show that the primers and gene targets are highly specific. No amplification of genes was observed in any of the off target body fluids, suggesting that both multiplexes are highly specific. Although the short amplicons were not used for sequencing, the Bioanalyzer results offer further support that the genes targeted in this study are highly specific to their respective body fluid, with no observed cross reactivity. Although no amplification of off-target body fluids was observed in the multiplex for long targets, there was some non-specific amplification of fragments nearly twice as large as expected fragments within target body fluids. Sequence data suggest that these longer, off target fragments are chimeric reads, as they appear to be a mixture

of two target genes within samples. Artificial chimeras can be generated by PCR due to incomplete primer extension, followed by amplification in later cycles.[55] Further optimization of the cycling parameters and cDNA input quantity for multiplex PCR may help decrease the introduction of chimeras.

Sequencing of long amplicons showed that the majority of reads were from the target body fluids. However, most samples had low levels of reads of off-target genes. It is possible that these reads are due to contamination, as all samples were amplified simultaneously. However, other studies that analyze mRNA targets through massively parallel sequencing (MPS) also had low levels of off-target reads in most samples. Both Hanson et al. and Ingold et al. observed read counts for off-target genes in most samples, but the contribution of these reads to the total number of reads was minimal.[54, 52] Hanson et al. converted read counts that comprised less than 0.5% of total reads to 0 for all analyses.[52] With the large amount of data gathered with high throughput sequencing, it appears that the occurrence of off-target gene reads is common in body fluid samples, and as long as off-target reads do not comprise a large percentage of the total number of reads for the sample, they may be ignored. However, the thresholds below which read counts can be ignored should be clearly defined and validated, particularly with more intensive sensitivity and mixture studies than what was explored in this study.

An unexpectedly high level of off-target read counts were obtained for some samples. In some saliva samples, there was higher than expected expression levels of semen genes, and higher than expected expression levels of semen and saliva genes in some blood samples. This appears to be correlated with reuse of flow cells. Saliva samples with a greater number of semen gene reads had the same barcode as a semen sample analyzed on the same flow cell in a previous run. Likewise, blood samples that had a greater number of semen or saliva gene reads had the

same barcode as a semen or saliva sample analyzed on the same flow cell. Although the MinION Wash Kit removes most of the library, there may be some residual library present after washing, and therefore ONT recommends barcoding libraries accordingly to eliminate contamination during data analysis.[56] The PCR Barcoding Kit I from ONT was used in this study, which only contains twelve unique barcodes. Due to a limited supply of flow cells, barcodes were repeated for samples in separate runs on the same flow cell. This can be avoided with the use of the PCR Barcoding Kit 96, which has 96 different barcodes and would therefore eliminate the need to repeat barcodes on a single flow cell.

There was also an unexpectedly high level of expression of off-target reads in vaginal fluid samples that could not be correlated to reuse of the flow cell. However, one sample contributed to 93.4% of the total number off-target reads for all vaginal fluid samples. The majority of the off-target reads were of HBA, followed by PRM2 and MMP7. It may be that the donor who contributed that vaginal fluid sample had just finished menstruating, which would account for the high number of blood and menstrual blood gene reads. The donor may have also engaged in vaginal sex within the seven days prior to collection, which would account for the high number of semen gene reads. This high level of expression of off-target reads was not observed in the other six vaginal fluid samples.

Housekeeping genes were not consistently observed across all sample types. The purpose of the housekeeping genes is to serve as an endogenous positive control and therefore should be observed uniformly in all samples that contain RNA. However, the expression of both housekeeping genes was very low in both semen and saliva samples, and only one housekeeping gene (B2M) was consistently expressed in blood samples. Conversely, housekeeping genes were highly expressed in vaginal fluid and menstrual blood samples. The total number of

housekeeping gene reads across all vaginal fluid samples was twice that of the total number of vaginal fluid gene reads. This difference in expression may be due to the collection method for vaginal fluid and menstrual blood swabs. Rather than pipetting liquid body fluids onto swabs, vaginal fluid samples and the menstrual blood sample were collected with a vaginal swab, which will result in collection of both body fluid and epithelial cells. The high expression level of housekeeping genes in these samples may be due to the presence of mRNA from epithelial cells.

Another study in which various housekeeping genes were assessed for their ability to serve as endogenous positive controls for forensic samples found that, although B2M was the most highly expressed gene in their gene panel, it was not expressed uniformly across all samples.[57] Authors of a recent study that explored mRNA analysis through MPS suggested that housekeeping genes should be left out of analysis, as the total number of reads of housekeeping genes far exceeded the total number of body fluid specific reads and therefore led to inefficiencies in sequencing of the body fluid specific genes.[52] In this study, the housekeeping genes were not informative for semen, saliva, or blood samples. Additionally, high expression levels of these genes in vaginal fluid and menstrual blood samples may have led to a decreased read count for the body fluid specific genes. In the future, it may be beneficial to remove housekeeping genes from analysis entirely.

In this study, only two genes were targeted for each body fluid. Addition of gene targets to the multiplex should be considered in order to increase the confidence in assignment of body fluid source. For each body fluid type, there was a predominant gene that comprised a majority of the reads for each sample. PRM2 was much more highly expressed than SEMG1, which poses an issue, as PRM2 will not be expressed in vasectomized individuals, while SEMG1 should be present in all seminal fluid. Other studies have shown that seminogelin genes are

highly and consistently expressed and have been able to be successfully targeted for identification of semen samples. [12, 53] Likewise, the saliva gene STATH and the blood gene ALAS2 have been shown to be highly and consistently expressed in saliva and blood samples, but were not expressed consistently in this study. [12, 40] Additionally, the vaginal fluid marker MYOZ1 and menstrual blood markers MMP7 and MMP10 did not perform as well in this multiplex, although they have been shown to be highly expressed in vaginal fluid and menstrual blood samples. [19, 58] This may be due to the length of amplicons targeted in this study. The MinION performs better with longer reads, with recommended target length of at least 300 bp.

Although some markers (PRM1, HTN3, HBA, and CYP2B7P1) performed well in the multiplex, it may be that the other mRNA markers degrade more quickly and therefore were not efficiently amplified in the multiplex targeting long fragments. One study that tested the performance of SEMG1 and PRM2 in 33-56 year old semen stains found that, although PRM2 was detected in up to 56 year old stains, SEMG1 was only detected in fresh semen samples. [59] This suggests that SEMG1 may degrade much more quickly than PRM2, which could account for the decreased number of reads. A study conducted by Zhao et al. (2016) found that while HBA was detected in nearly 90% of 50 year old blood samples, ALAS2 was detected in less than 3% of 30 year old blood samples and in 0% of 50 year old samples, suggesting that this marker, like SEMG1, degrades more quickly. [60] Additionally, both Setzer et al. (2008) and Sirker et al. (2016) found that saliva markers (particularly STATH) are more subject to hydrolytic damage and are therefore observed less often and more sporadically in older or poorly stored samples. [4, 6] The occurrence of decreased or sporadically observed reads of body fluid genes in this study may be due to degradation of mRNA fragments in the samples. Although the primers were designed to target the innermost portions of mRNA fragments, the efficiency of amplification

may be decreased due to shortening of the mRNA fragments. The read length requirement for acquiring high quality data with the MinION would likely not be ideal for highly degraded samples, and other markers may be explored that are less subject to degradation for these body fluids. This is supported by the amplicon length data from the samples amplified in the multiplex targeting shorter amplicons, as amplification of these genes appeared to be slightly more successful with shorter targets.

Menstrual samples were ordered from Lee Biosolutions and analyzed in the multiplex targeting short amplicons. However, no amplification of blood, vaginal fluid, menstrual blood, or housekeeping genes was apparent according to Bioanalyzer results. Because of this, these samples were removed from the analysis. The storage of these samples prior to receipt is unknown, so it may be that the samples were improperly stored and any mRNA was too degraded for analysis. A menstrual blood sample collected by donors via vaginal swab showed amplification of vaginal fluid, blood, menstrual blood, and housekeeping genes with both multiplexes. However, no apparent amplification or reads of the menstrual blood gene MMP10 was observed, which has been shown to be one of the more highly expressed menstrual blood biomarkers.[19] Failure of MMP10 amplification could be due to improper primer design or due to storage conditions. More menstrual blood samples should be added for analysis to further address this issue, as MMP10 should be present in menstrual blood samples. However, a mixture of blood, vaginal fluid, and menstrual blood genes suggests that the body fluid source may be menstrual blood, which should be taken into account during analysis.

Although decreased read counts for some genes may be due to degradation of the mRNA, it should also be considered that some genes are simply more highly expressed than others. Adding genes to the multiplex would allow forensic analysts to see a general pattern of gene

expression, which could be a better analysis method than just looking for the presence or absence of markers. Hanson et al. (2018) suggest that observing the general trend of expression of genes is informative to the quantity and quality of the mRNA within the sample and therefore this trend may be a valuable analysis tool. [52]

Marker sensitivity

Semen genes were detected down to 1:100 dilutions of semen, but not in 1:1,000 dilutions. The saliva gene HTN3 was only detected in 1:10 dilutions of saliva, but STATH was not detected in any of the dilution samples. The blood gene HBA was detected in down to 1:1,000 dilutions of blood, but ALAS2 was not detected in most samples. Another high throughput sequencing study showed that saliva and vaginal fluid genes were not detectable below 10 ng of RNA for sequencing, while semen and blood genes were detectable down to 5 ng of total input. [52] In order to fully assess the sensitivity of this assay, another RNA quantification method may need to be implemented to determine the amount of RNA in samples prior to cDNA synthesis and amplification, and how varying RNA inputs affect the total read count of body fluid specific genes.

Semen and vaginal fluid were both detectable in mixture samples, demonstrating the ability of multiple body fluids to be detected in two person mixtures with this method. Other mixture combinations should be added in the future, as the ability to detect mixtures of biological fluids is important in forensic casework. The overall sequence yield for the vaginal fluid/semen mixture samples appeared to be negatively affected by reuse of the flow cell, as the total sequence yield was much less than that of other vaginal swabs analyzed in previous runs. It is likely that a higher number of reads could be achieved by running these samples again on a flow cell with a greater number of available pores.

MinION performance

Despite the 1 µg of DNA input recommended for the Ligation Sequencing Kit 1D, the MinION appeared to be able to generate over 1,000X coverage for most body fluid genes. Forensic samples, and particularly mRNA samples, are generally of lower quantity and quality than what is recommended for the MinION. Additionally, the MinION has largely been adapted for sequencing of very large fragments, which is something that is not compatible with mRNA analysis in forensic samples. Despite this, the MinION was able to successfully sequence amplicons 344-440 bp in length generated from low-level samples. Additionally, the accuracy of long read sequencing platforms like the MinION have been known to be lower than that of sequencing platforms geared towards sequencing shorter fragments, like the Illumina® MiSeq™. However, updates to flow cell chemistry and analysis software have increased the overall accuracy of basecalling for the MinION platform.[61] In September of 2017, the Albacore v2.0.1 software was released, which uses the raw data generated by the MinION for basecalling. This was shown to increase the overall accuracy of reads.[61] Improvements to sequencing kit and flow cell chemistry and the analysis software algorithms have expanded the reach of this platform. This proof of concept study demonstrates that the MinION sequencer is capable of successfully sequencing and identifying mRNA amplicons for body fluid identification in forensic type samples.

Although the MinION was able to generate high quality sequence data from these samples, optimization and further validation of this sequencing method are necessary. One issue that should be addressed in the future is the normalization of barcoded amplicons prior to sequencing preparation. In the end-prepping portion of MinION sequencing workflow with the Ligation Sequencing Kit 1D, all barcoded amplicons are pooled to achieve a 45 µl total input

volume. As twelve samples were sequenced at a time, 3.75 µl of each sample were combined to achieve the 45 µl input volume. However, after barcoding, amplicon concentration varied from sample to sample. Normalizing samples according to amplicon concentration may allow for more consistent total read counts for each sample on the flow cell, as the total number of read counts in this study was highly variable from sample to sample within a sequencing run. Additionally, the MinION protocol recommends using 0.2 pmoles of end-prepped DNA for adapter ligation. However, when 0.2 pmoles were added for adapter ligation, the read counts for all samples were very low, with a total read count of 10,103 for eleven samples. Adding the full volume of end-prepped DNA (30 µl) without normalization led to acquisition of 10-60 times more reads of body fluid genes.

Although ONT has developed some low input sequencing kits and protocols during the course of this study, they are only suitable for gDNA as input. A few kits have been developed that require RNA or cDNA as input for the sequencing workflow. One that would be applicable to the pre-sequencing workflow designed in this study is the Sequence-specific cDNA-PCR sequencing using the cDNA-PCR Sequencing Kits (SQK-PCS108). This protocol uses the strand-switching method of cDNA synthesis with user designed primers to allow for targeted cDNA synthesis, followed by PCR of full length transcripts and rapid adapter ligation.[62] This method could potentially decrease the total workflow time. Additionally, the workflow is recommended for low RNA input. However, it appears to select full length cDNA transcripts, which may not be achievable with forensic samples. Additionally, the kit is not compatible with any barcoding kits at the moment, although it is likely that barcoding will be added in the future.

Use of the strand-switching method of cDNA synthesis may increase sequencing yield. Strand-switching cDNA synthesis uses the Moloney murine leukemia virus (MMLV) reverse

transcriptase to add additional nucleotides as it reaches the 5' end of mRNA to act as a binding site for the template switching (TS) oligo.[63] Binding of the TS oligo allows MMLV to switch strands and continue cDNA synthesis to complete the 5' end of the transcript, which has been shown to increase overall cDNA production.[63] Additionally, optimization of the 1D PCR barcoding protocol should be explored in order to increase sequencing yield. Wei et al. demonstrated that, with the Genomic Sequencing Kit protocol, decreasing the volume of the adapter ligation reaction increased production of adapter ligated sequences by 48%, and adding a 1-2 hour incubation at 4°C after adapter ligation increased products by 61-63%.[64] Decreasing the adapter ligation reaction volume would extend the use of the sequencing kit and therefore decrease the cost per sample, as the adapter ligation mix is often the first reagent in the kit to be used up.

Flow cells appear to experience a decrease in yield after reuse. The occupancy of a flow cell examined in this study decreased drastically after three consecutive runs, which led to decreased overall sequence yield. Because the flow cells are currently quite expensive, this poses a problem. ONT has two options for washing flow cells. Both methods begin with addition of Solution A in the Wash Kit, which releases DNA strands from the nanopores in the membrane.[65] After loading Solution A, Solution B can be added to remove the library and prime the flow cell for loading of the next library immediately, or Storage Buffer can be added if the flow cell will be stored.[65] If Solution B is added, a platform QC cannot be completed to assess the number of single active pores available for sequencing. Platform QC of flow cells with Storage Buffer must be completed prior to adding the next library. With the three consecutive runs on the single flow cell analyzed in this study, Solution B was added for washing and the next library was added immediately without doing a platform QC on the flow cell. In the future,

it may be wise to alternate flow cells and add Storage Buffer for flow cell washing, rather than Solution B. This would allow for a platform QC to be completed on a flow cell prior to every run in order to ensure enough pores are available for efficient sequencing of libraries.

The cost per sample in this study, with reuse of flow cells 2-4 times, was $60.21-$78.96. The cost per sample could be brought down with further optimization of the sequencing workflow. In this study, 12 barcoded samples were combined into each sequencing reaction. However, with the 96 barcode kit, more samples could be pooled into one sequencing reaction. Additionally, if flow cells are ordered in bulk, the price of individual flow cells decreases from $900 to $475. Flow cells are recommended to be used within 8 weeks of receipt, so ordering of flow cells in bulk would be dependent on the throughput of individual crime labs. Reuse of flow cells decreases the number of active pores, which would decrease sequence yield for subsequent runs. Oxford Nanopore Technologies recommends only using flow cells with at least 800 single active pores, which is determined by completing a platform QC with MinKNOW prior to loading the library. Some flow cells maintained 800 active pores after 2 to 3 uses, while the pore count for other flow cells fell below 800 pores after only one use. Flow cells also appear to lose more single active pores after longer sequencing runs. Because of this, sequencing runs were stopped after 4-5 hours in this study. Allowing sequencing runs to continue for longer may increase the amount of sequence data generated, so the optimal run time for these types of samples should be determined in order to maximize data acquisition while also minimizing the cost per sample.

The MinION platform should also be compared to other sequencing platforms, like the Illumina® platform. The MinION method is appealing due to its small size, simple setup, simple library preparation, and real-time data analysis. The library preparation for the MinION is quick and simple, so it would be easy to train analysts to use this method. The cloud-based EPI2ME

software allows the user to monitor the number, size, and quality of reads generated during a sequencing run from anywhere, allowing the user to stop the sequencing run once a sufficient amount of data is obtained and load the next library. Additionally, the price of the sequencer is only $1,000, although the reagents and flow cells are still expensive. However, there are some drawbacks to using the MinION. The longevity of the flow cells appears to be unpredictable, which could cost crime labs more money than anticipated. The MinION also requires longer fragments of DNA/RNA in order to be able to obtain high quality sequences, which is not ideal for degraded forensic samples. Conversely, the Illumina® platform offers the potential to target much smaller fragments and more options for low quantity and quality samples. Other sequencing platforms should be explored with this body fluid identification method in order to compare the overall quantity and quality of data generated to that obtained with the MinION. The capability of other platforms to be integrated into the workflow of crime labs should also be assessed in comparison to that of the MinION.

Conclusions

This proof of concept study explored the ability of the MinION sequencer to generate high quality sequences from forensic mRNA samples for body fluid identification purposes. A DNA/RNA co-extraction method allowed for the generation of DNA and RNA profiles from single half swabs with semen, saliva, blood, vaginal fluid, and menstrual blood. A multiplex PCR designed to target two body fluid specific genes per body fluid type and two housekeeping genes successfully amplified cDNA in body fluid samples with little to no cross reactivity. By targeting longer mRNA fragments (>300 bp) through PCR, and with preparation with the 1D Ligation Sequencing Kit by ONT, high quality sequences were mapped to reference genes with >1,000X coverage of expected gene in most samples and even in diluted samples and mixtures.

Although the method in this study was successfully used to identify body fluid specific mRNAs, further optimization is needed to increase sequence yield. Additionally, the ability of this sequencer to generate high quality data from more degraded samples should be assessed, since the length of the fragments targeted in this study may not be ideal in more degraded, older samples. Comparison of this platform to other high throughput sequencing platforms in the future would give the forensic community better insight into the benefits and drawbacks of introducing these different body fluid identification methods into crime labs. Overall, the MinION appears to be capable of generating sequences with much smaller input quantities than recommended when PCR is integrated into the workflow, allowing it to be adapted for use with forensic type samples.

REFERENCES

1. Juusola J, Ballantyne J. *Messenger RNA profiling: a prototype method to supplant conventional methods for body fluid identification*. Vol 135: Forensic Science International; 2003.
2. Alberts B, Johnson A, Lewis J. *Molecular biology of the cell [Internet]*. 4th edition ed: Garland Science.
3. Sijen T. Molecular approaches for forensic cell type identification: On mRNA, miRNA, DNA methylation and microbial markers. *Forensic Science International: Genetics*. 2015;18:21-32.
4. Setzer M, Juusola J, Ballantyne J. Recovery and Stability of RNA in Vaginal Swabs and Blood, Semen, and Saliva Stains. *Journal of Forensic Science*. 2008;53(2):296-305.
5. Zhao H, Wang C, Yao L, et al. Identification of aged bloodstains through mRNA profiling: Experiments results on selected markers of 30- and 50-year-old samples. *Forensic Science International*. 2016.
6. Sirker M, Schneider PM, Gomes I. A 17-month time course study of human RNA and DNA degradation in body fluids under dry and humid environmental conditions. *International Journal of Legal Medicine*. 2016;130:1431-1438.
7. Grabmuller M, Medea B, Courts C. Comparative evaluation of different extraction and quantification methods for forensic RNA analysis. *Forensic Science International: Genetics*. 2015;16:195-202.
8. Schweighardt AJ, Tate CM, Scott KA, Harper KA, Robertson JM. Evaluation of commercial kits for dual extraction of DNA and RNA from human body fluids. *Journal of Forensic Sciences*. 2015;60(1):157-165.
9. Watanabe K, Iwashima Y, Akutsu T, Sekiguchi K, Sakurada K. Evaluation of a co-extraction method for real-time PCR identification and DNA typing. *Legal Medicine*. 2014;16:56-59.
10. Jakubowska J, Maciejewska A, Bielawski KP, Pawlowski R. mRNA heptaplex protocol for distinguishing between menstrual and peripheral blood. 2014;13:53-60.
11. Lindenbergh A, Pagter M, Ramdayal G, et al. A multiplex (m)RNA-profiling system for the forensic identification of body fluids and contact traces. *Forensic Science International: Genetics*. 2012;6:565-577.
12. Haas C, Hanson E, Anjos MJ, et al. RNA/DNA co-analysis from human saliva and semen stains--Results of a third collaborative EDNAP exercise. *Forensic Science International: Genetics*. 2013;7:230-239.
13. Zubakov D, Kokmeijer I, Ralf A, et al. Towards simultaneous individual and tissue identification: A proof-of-principle study on parallel sequencing of STRs, amelogenin, and mRNAs with the Ion Torrent PGM. *Forensic Science International: Genetics*. 2015;15:122-128.

14. Chong KW, Wong Y, Ng BK, Thong Z, Syn CK. Development of a RNA profiling assay for biological tissue and body fluid identification. *Forensic Science International: Genetics*. 2015;5:e197-e198.
15. Xu Y, Xie J, Cao Y, et al. Development of highly sensitive and specific mRNA multiple system (XCYR1) for forensic human body fluids and tissues identification. *PLoS ONE*. 2014:e100123.
16. Lilja H, Abrahamson PA, Lundwall A. Seminogelin, the predominant protein in human semen. *The Journal of Biological Chemistry*. 1989;264(3):1894-1900.
17. Wykes SM, Krawetz SA. The structural organization of sperm chromatin. *Journal of Biological Chemistry*. 2003;278(32):29471-29477.
18. Marbaix E. Role of matrix metalloproteinases in the endometrium in normal and pathologic menstruation. *Bull Mem Acad Med Belg*. 2005;160(5):232-238.
19. Haas C, Hanson E, Anjos MJ, et al. RNA/DNA co-analysis from human menstrual blood and vaginal secretion stains--Results of a fourth and fifth collaborative EDNAP exercise. *Forensic Science International: Genetics*. 2014;8:203-212.
20. Hanson E, Ingold S, Haas C, Ballantyne J. Targeted multiplexed next generation RNA sequencing assay for tissue source determination of forensic samples. *Forensic Science International: Genetics*. 2015;5:e441-e443.
21. Hunter GA, Ferreira GC. 5-Aminolevulinate synthase: catalysis of the first step of heme biosynthesis. *Cell Mol Biol*. 2009;55(1):102-110.
22. Bunn F. Subunit assembly of hemoglobin: an important determinant of hematologic phenotype. *Blood*. 1987;69(1):1-6.
23. Imamura Y, Wang PL, Masuno K, Sogawa N. Salivary protein histatin 3 regulates cell proliferation by enhancing p57Kip1 and heat shock cognate protein 70 ubiquitination. *Biochemical and Biophysical Research Communications*. 2016;470:269-274.
24. Weidner T, Dubey M, Breen NF, et al. Direct observation of phenylalanine orientations in statherin bound to hydroxyapatite surfaces. *J Am Chem Soc*. 2012;134(21):8750-8753.
25. Takada F, Vander Woude DC, Tong HQ, et al. Myozenin: An α-actin-and-γ-filamin-binding protein of skeletal muscle Z lines. *PNAS*. 2001;98(4):1595-1600.
26. National Center for Biotechnology Information. Homo Sapiens complex locus CYP2B7P1, endcoding cytochrome p450, family 2, subfamily B, polypeptide 7, pseudogene 1. [cited 2017 June 19]..
27. National Center for Biotechnology Information. G6PDH glucose-6-phosphate dehydrogenase [Homo sapiens (human)].. 2017.
28. Laver T, Harrison J, O'Neill PA, et al. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*. 2015;3:1-8.
29. Lowman NJ, Watson M. Successful test launch for nanopore sequencing. *Nature Methods*. 2015;12(3):303-304.
30. Oxford Nanopore Technologies. Chemistry Technical Document: Ligation Sequencing Kit family. *community.nanoporetech.com*. https://community.nanoporetech.com/technical_documents/chemistry-technical-document/v/chtd_500_v1_revm_07jul2016/ligation-sequencing-kit-family . Accessed May 17, 2018.

31. Life Technologies. PrepFiler and Prepfiler BTA Forensic DNA Extraction Kits. 2012. https://www3.appliedbiosystems.com/cms/groups/applied_markets_support/documents/generaldocuments/cms_096102.pdf . Accessed April 28, 2017.
32. Qiagen. RNeasy Mini Handbook. 2012. http://www.bea.ki.se/documents/EN-RNeasy%20handbook.pdf. Accessed April 28, 2017.
33. Life Technologies. Quantifiler HP and Trio DNA Quantification Kits User Guide. 2015. https://tools.thermofisher.com/content/sfs/manuals/4485354.pdf . Accessed April 28, 2017.
34. Agilent Technologies. Agilent RNA 6000 Pico Kit Guide. 2013. https://www.agilent.com/cs/library/usermanuals/Public/G2938-90046_RNA600Pico_KG_EN.pdf . Accessed May 1, 2017.
35. Schroeder A, Mueller O, Stocker S, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology*. 2006;7(3):doi:10.1186/1471-2199-7-3.
36. Life Technologies. TURBO DNA-free Kit User Guide. 2012. https://tools.thermofisher.com/content/sfs/manuals/cms_055740.pdf . Accessed May 3, 2017.
37. New England BioLabs Inc. ProtoScript First Strand cDNA Synthesis Kit Instruction Manual. 2014. https://www.neb.com/~/media/Catalog/All-Products/8DC78EF3331D476F9C263B572910651B/Datacards%20or%20Manuals/manualE6300.pdf . Accessed May 3, 2017.
38. Life Technologies. GlobalFiler PCR Amplification Kit. https://tools.lifetechnologies.com/content/sfs/manuals/4477604.pdf . Accessed May 3, 2017.
39. Haas C, Klesser B, Maake C, Bar W, Kratzer A. mRNA profiling for body fluid identification by reverse transcription endpoint PCR and realtime PCR. *Forensic Science International: Genetics*. 2009;3:80-88.
40. Haas C, Hanson E, Kratzer A, Bar W, Ballantyne J. Selection of highly specific and sensitive mRNA biomarkers for the identification of blood. *Forensic Science International: Genetics*. 2011;5:449-458.
41. Berge M, Bhoelai B, Harteveld J, Matai A, Sijen T. Advancing forensic RNA typing: On non-target secretions, a nasal mucosa marker, a differential co-extraction protocol and the sensitivity of DNA and RNA profiling. *Forensic Science International: Genetics*. 2016;20:119-129.
42. Berge M, Carracedo A, Gomes I, et al. A collarborative European exercise on mRNA-based body fluid/skin typing and interpretation of DNA and RNA results. *Forensic Science International: Genetics*. 2014;10:40-48.
43. Oxford Nanopore Technologies. 1D PCR Barcoding Amplicons (SQK-LSK108) Protocol. https://community.nanoporetech.com/protocols/pcr-barcoding-amplicons-r/v/pbae_9008_v108_revr_18oct2016/pcr-barcoding-amplicons . Accessed February 20, 2018.
44. Qiagen. Multiplex PCR Handbook. 2010. https://www.qiagen.com/it/resources/resourcedetail?id=a541a49c-cd06-40ca-b1d2-563d0324ad6c&lang=en . Accessed May 3, 2017.

45. Haas C, Hanson E, Banemann R, et al. RNA/DNA co-analysis from human skin and contact traces – results of a sixth collaborative EDNAP exercise. *Forensic Science International: Genetics*. 2015;16:139-147.
46. Whitehead Institute for Biomedical Research. IDT's PrimerQuest Tool
47. Agilent Technologies. Agilent DNA 1000 Kit Guide. 2016. https://www.agilent.com/cs/library/usermanuals/Public/G2938-90014_DNA1000Assay_KG.pdf . Accessed February 20, 2018.
48. Applied Biosystems. ExoSAP-IT PCR Product Cleanup. 2017. https://tools.thermofisher.com/content/sfs/manuals/78200b.pdf . Accessed February 20, 2018.
49. Takara Bio Inc. TaKaRa Taq DNA Polymerase and Premix Taq DNA Polymerase. http://www.clontech.com/US/Products/PCR/General_Purpose/Taq_and_Premix_Taq?sitex=10020:22372:US . Accessed March 18, 2018.
50. Agencourt Bioscience Corporation. Agencourt AMPure XP PCR Purification. https://www.beckmancoulter.com/wsrportal/bibliography?docname=Protocol_000387v001.pdf . Accessed March 1, 2018.
51. Loman N, Quinlan A. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics*. 2014.
52. Haas C, Mulheim C, Kratzer A, Bar W, Maake C. mRNA profiling for the identification of sperm and seminal plasma. *Forensic Science International: Genetics Supplement Series*. 2009:534-535.
53. Ingold S, Dorum G, Hanson E, et al. Body fluid identification using a targeted mRNA massively parallel sequencing approach--results of a EUROFORGEN/EDNAP collaborative exercise. *Forensic Science International: Genetics*. 2018;34:105-115.
54. Smyth RP, Schlub TE, Grimm A, et al. Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene*. 2010;469(1-2):45-51.
55. Hanson E, Ingold S, Haas C, Ballantyne J. Messenger RNA biomarker signatures for forensic body fluid identification revealed by targeted RNA sequencing. *Forensic Science International: Genetics*. 2018;34:206-221.
56. Technologies ON. Washing MinION Flow Cells. *nanoporetech.com*. https://community.nanoporetech.com/protocols/washing-minion-flow-cells/v/wke_1012_v1_revk_08apr2016 . Accessed April 15, 2018.
57. Moreno LI, Tate CM, Knott EL, et al. Determination of an effective housekeeping gene for the quantification of mRNA for forensic applications. *Journal of Forensic Sciences*. 2012;57(4):1051-1058.
58. Hanson E, Ballantyne J. Highly specific mRNA biomarkers for the identification of vaginal secretions in sexual assault investigations. *Science and Justice*. 2013;53:14-22.
59. Nakanishi H, Hara M, Takahashi S, Takada A, Saito K. Evaluation of forensic examination of extremely aged seminal stains. *Legal Medicine*. 2014;16:303-307.
60. Zhao H, Wang C, Yao L, et al. Identification of aged bloodstains through mRNA profiling: experiments results on selected markers of 30- and 50-year-old samples. *Forensic Science International*. 2017;272:e1-e6.

61. Oxford Nanopore Technologies. New basecaller now performs 'raw basecalling', for improved sequencing accuracy. https://nanoporetech.com/about-us/news/new-basecaller-now-performs-raw-basecalling-improved-sequencing-accuracy . Accessed April 16, 2018.
62. Oxford Nanopore Technologies. Sequence-specific cDNA-PCR sequencing protocol. *community.nanoporetech.com*. https://community.nanoporetech.com/protocols/sequence-specific-cdna-pcr-sequencing/v/cpss_9060_v108_reva_15mar2018 . Accessed May 17, 2018.
63. Wang B. Use of template switching oligos (TS oligos, TSOs) for efficient cDNA library construction. *IDT Integrated DNA Technologies*. 2018. https://www.idtdna.com/pages/education/decoded/article/use-of-template-switching-oligos-(ts-oligos-tsos)-for-efficient-cdna-library-construction . Accessed May 17, 2018.
64. Wei S, Williams Z. Rapid Short-Read Sequencing and Aneuploidy Detection Using MinION Nanopore Technology. *Genetics*. 2016;202(1):37-44.
65. Oxford Nanopore Technologies. Washing MinION Flow Cells Protocol. *community.nanoporetech.com*. https://community.nanoporetech.com/protocols/washing-minion-flow-cells/v/wke_1012_v1_revk_08apr2016 . Accessed April 20, 2018.

APPENDIX

Table A1. DNA quantity and the completeness of profile based on autosomal STR loci for all body fluid samples.

Body Fluid	Replicate	DNA Quantity (ng)	Number of Observed Alleles	Profile Completeness (%)
Semen	1	0.29	42/42	100.00
	2	0.68	42/42	100.00
	3	0.13	42/42	100.00
	4	0.03	42/42	100.00
	5	0.56	42/42	100.00
	6	0.40	42/42	100.00
	7	0.21	42/42	100.00
	8	0.73	42/42	100.00
Saliva	1	0.00	21/42	50.000
	2	0.00	33/42	78.570
	3	0.06	42/42	100.00
	4	0.37	42/42	100.00
	5	0.12	42/42	100.00
	6	0.04	42/42	100.00
	7	0.06	42/42	100.00
	8	0.01	42/42	100.00
Blood	1	0.00	39/42	92.860
	2	0.04	42/42	100.00
	3	0.02	42/42	100.00
	4	0.02	42/42	100.00
	5	0.03	42/42	100.00
	6	0.02	42/42	100.00
	7	0.07	42/42	100.00
	8	0.02	42/42	100.00
Vaginal Fluid	1	0.04	38/42	90.480
	2	4.10	42/42	100.00
	3	2.78	42/42	100.00
	4	1.01	42/42	100.00
	5	3.99	42/42	100.00
	6	3.22	42/42	100.00
	7	7.40	42/42	100.00
	8	8.69	42/42	100.00
Menstrual blood	1	0.22	42/42	100.00
	2	0.43	42/42	100.00
	3	0.05	42/42	100.00
	4	0.22	42/42	100.00
	5	0.21	42/42	100.00
	6	0.43	42/42	100.00
	7	0.10	42/42	100.00
	8	0.49	42/42	100.00

Table A2. DNA quantity and the completeness of profile based on autosomal STR loci for 3 replicates of 1:10, 1:100, and 1:1,000 dilutions of semen, saliva, and blood.

Body Fluid	Dilution Factor	DNA Quantity (ng)	Number of Observed Alleles	Profile Completeness (%)
Semen	1:10	0.01	42/42	100.0
		0.03	42/42	100.0
		0.08	42/42	100.0
	1:100	0.00	9/42	21.43
		0.00	11/42	26.19
		0.00	39/42	92.86
	1:1,000	0.00	2/42	04.76
		0.00	0/42	00.00
		0.00	7/42	16.67
Saliva	1:10	0.00	11/42	26.19
		0.01	42/42	100.0
		0.01	42/42	100.0
	1:100	0.00	14/42	33.33
		0.00	7/42	16.67
		0.00	0/42	00.00
	1:1,000	0.00	0/42	00.00
		0.00	0/42	00.00
		0.00	0/42	00.00
Blood	1:10	0.00	35/42	83.33
		0.00	38/42	90.48
		0.00	39/42	92.86
	1:100	0.00	8/42	19.05
		0.00	40/42	95.24
		0.00	2/42	04.76
	1:1,000	0.00	0/42	00.00
		0.00	1/42	02.38
		0.00	0/42	00.00

Table A3. DNA quantity and the completeness of profile based on autosomal STR loci for vaginal swabs with 7.5 µl, 15 µl, and 30 µl of semen for 3 sets of donors.

Semen Volume (μl)	DNA Quantity (ng)	Contributer	Number of Observed Alleles	Profile Completeness (%)
7.5	7.15	Major	42/42	100.0
		Minor	28/42	66.67
	4.72	Major	42/42	100.0
		Minor	26/42	61.90
	14.62	Major	42/42	100.0
		Minor	19/42	45.24
15	7.03	Major	42/42	100.0
		Minor	42/42	100.0
	12.61	Major	42/42	100.0
		Minor	41/42	97.62
	20.23	Major	42/42	100.0
		Minor	38/42	90.48
30	8.63	Major	42/42	100.0
		Minor	42-42	100.0
	11.47	Major	42/42	100.0
		Minor	2/42	100.0
	10.05	Major	42/42	100.0
		Minor	40/42	95.24

Table A4. Total number of reads for each body fluid sample analyzed in this study. Numbers highlighted in green comprise >10% of total reads for the sample, while numbers highlighted in yellow represent <5% of total reads and numbers in while comprise <1% of all reads.

Body Fluid	Donor	SEMG1	PRM2	HTN3	STATH	HBA	ALAS2	CYP2B7P1	MYOZ1	MMP7	MMP10	B2M	UBC	Total
Semen	1	4,504	20,410	77	43	8	4	33	0	0	0	48	2670	27,797
	2	4,504	17,016	13	0	0	0	0	0	0	0	8	3	21,544
	3	44	4,717	12	0	0	0	0	0	0	0	7	8	4,788
	4	55	9,248	17	0	0	0	1	0	0	0	6	6	9,333
	5	17,770	51,754	27	0	0	0	0	0	0	0	7	3	69,561
	6	48	18,067	15	0	0	0	0	0	0	0	5	5	18,140
	7	108	40,057	34	0	0	0	0	0	0	0	11	7	40,217
	8	19,158	53,692	21	0	0	0	0	0	0	0	2,289	2,994	78,154
Saliva	1	32	174	3,767	0	0	0	0	0	0	0	2	2	3,977
	3	34	117	6,371	0	0	0	0	0	0	0	1,159	4	7,685
	4	65	291	2,878	0	0	0	0	0	0	0	25	5	3,264
	5	20	68	152,750	90,927	87	16	380	1	13	0	24,195	380	268,837
	6	390	1,601	3	1,526	25	3	8	0	0	0	7	0	3,563
	8	6	549	6,867	3,514	95	14	18	0	0	0	1,245	1,728	14,036
Blood	1	0	6	16	5	3,466	1,693	10	0	0	0	1937	137	7,270
	2	1,872	5,848	11	4	38,581	8	14	0	0	0	1,521	15	47,874
	3	5	1,840	4	4	8,035	982	2	0	0	0	2,757	119	13,748
	4	9	4,453	14	8	27,718	2,422	15	0	0	0	9,828	582	45,049
	5	1,618	5,102	6	1	12,616	1,004	14	0	0	0	5,277	239	25,877
	6	5	25	439	4	17,583	462	12	0	0	0	4,437	1,048	24,015
	7	1	13	4	0	1,875	635	1	0	0	0	793	116	3,438
	8	4	14	545	4	5,626	2,194	5	0	0	0	2,943	277	11,612
Vaginal Fluid	1	8	43	435	9	138	12	17,219	0	0	0	53	5,722	23,639
	2	19	64	370	212	92	22	155,684	539	8	0	56,655	106,395	320,060
	3	14	19	0	0	8	0	3,670	1	1	0	117	8,084	11,914
	5	13	9	0	0	11	0	21,752	416	0	0	4,300	15,458	41,959
	6	29	18	0	0	23	0	124	0	3	0	222	73,069	73,488
	7	34	47	0	0	23	0	12,418	1	3	0	7,512	49,847	69,885
	8	2,045	74	0	0	20,055	0	63,732	1	1,646	0	161,368	90,555	339,476
Menstrual blood	1	22	55	117	66	35192	5815	8485	2	2925	0	31004	17120	100,803

Table A5. Total number of reads for each sample in the vaginal fluid/semen mixture samples. Numbers highlighted in green comprise >10% of total reads for the sample, while numbers highlighted in yellow represent <5% of total reads and numbers in while comprise <1% of all reads.

Volume of Semen (μ)l	SEMG1	PRM2	HTN3	STATH	HBA	ALAS2	CYP2B7P1	MYOZ1	MMP7	MMP10	B2M	UBC	Total
7.5	65	344	0	0	551	0	1,774	30	0	0	967	1,360	5,091
	1,381	1,030	0	0	3,833	0	9,402	37	108	0	12,519	9,544	37,854
	460	165	0	0	1,218	25	187	0	0	0	1,033	352	3,440
15	455	1,107	0	0	10	0	1,768	1	0	0	1,345	1,426	6,112
	926	684	0	0	8	0	3,007	0	0	0	1,296	5,237	11,158
	4,324	1,102	0	0	837	0	427	0	0	0	4,130	1,820	12,640
30	322	525	0	0	4	0	907	1	0	0	489	1,368	3,616
	2,174	1,492	0	0	2,194	0	4,021	2	0	0	1,512	9,985	21,380
	1,380	1,744	0	0	844	0	756	0	2	0	1,801	803	7,330